UCLA Biostatistics Topics
200ABC, 202AB

Jonathan Hori

Fall 2022 - Spring 2023

This document contains a list of basically every topic covered in the first year of the UCLA Biostatistics MS and PhD core sequence.

Document as of September 25, 2023.

Contents

1 Biostat 200A

- Descriptive vs. inferential statistics
- Population + parameters vs. sample + statistics
- Quantitative vs. qualitative data. Nominal, ordinal, interval, ratio data
- Visual displays:
  - stemplots (stem + leaf)
  - boxplots (box + whisker), whisker length = 1.5 * IQR
  - bar chart
  - pie chart
  - frequency table. Columns = interval, implied limits, midpoint, freq, relative %, cumulative %
  - frequency polygon, cumulative distribution polygon
  - histogram
  - two-way frequency tables (contingency tables)
  - scatterplots (with/without jitter)
- Percentiles: (k+1)st point if $np/100 \notin \mathbb{Z}, k = \lfloor np/100 \rfloor$, otherwise average of obs (np/100) and (np100 + 1)
- Numerical summary measures - Measures of location: mode, mean, median. Possible number of each per dataset. Relationships for skewed data.
- Other "averages": midrange = $(X_{(n)} + X_{(1)})/2$, trimmed mean = $(\sum_{x \neq min,max} x)/(n-2)$, geometric mean = $(\prod_{i=1}^{n} x_i)^{1/n}$, harmonic mean = $\frac{1}{1/n \sum_{i=1}^{n} 1/x_i}$
- Numerical summary measures - Measures of variation: range, IQR, variance, standard deviation
- Sample moments: $m'_r = 1/n \sum_{i=1}^{n} x_i^r$ (about 0). Relationship to mean, variance, skewness, kurtosis.
- Grouped data: k groups, midpoints $x_{i,mid}$, frequencies f_i:
  - $\bar{X} = \frac{\sum f_i x_{i,mid}}{n}$
  - $s = \sqrt{\frac{\sum f_i x_{i,mid}^2 - (\sum f_i x_{i,mid})^2/n}{n-1}}$
  - percentiles $p = l + [\frac{(\%/100)*n-c}{f}] * w$
- Effects of uniform changes to data on mean and std: $x_i \to cx_i + b$
- Chebyshev's inequality: $P(g(X) \geq k) \leq \frac{E[g(X)]}{k} \forall k > 0$
- Definitions of probability: classical vs. relative frequency vs. subjective/personal vs. axiomatic
- $P(A) + P(A^C) = 1$
- Adding probabilities: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. The intersection has probability 0 if A and B are mutually exclusive.

- Conditional probability and independence:
  - $P(B|A) = \frac{P(A \cap B)}{P(A)}$
  - $P(A \cap B) = P(A)P(B|A)$
  - A, B independent iff $P(A \cap B) = P(A)P(B)$
- Bayes theorem: $P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A)+P(B|A^C)P(A^C)}$.
  - Rewriting as posterior odds = prior odds × LR:

$$\frac{P(A|B)}{P(A^C|B)} = \frac{P(A)}{P(A^C)} \times \frac{P(B|A)}{P(B|A^C)}$$

- Odds = $\frac{P(B|A)}{P(B^C|A)}$
- Relative risk: $RR = \frac{P(B|A)}{P(B|A^C)}$
- Diagnostic screening (using 2x2 table):
  - Percent correctly classified = total correctly classified / total
  - Sensitivity = true-positive rate = $P(test+|disease+)$
  - Specificity = true-negative rate = $P(test-|disease-)$
  - Positive predictive value = Proportion of those testing positive having disease = $P(disease+|test+)$. Depends on disease prevalence through prior.
  - Negative predictive value = Proportion of those testing negative not having disease = $P(disease-|test-)$. Depends on disease prevalence through prior.
- Sampling schemes: chance sampling, simple random sample, general probability sampling, cluster sampling, stratified sampling, systematic sampling, convenience sampling
- Bias: selection bias, nonresponse bias, unintentional bias
- Definition of random variable. Discrete vs. continuous. Cumulative distribution functions.
- Discrete distributions:
  - binomial
  - hypergeometric
  - poisson
- Continuous distributions:
  - Normal (how to standardize)
  - Chi-squared
- Population moments: $E(X) = \int_{-\infty}^{\infty} xf(x)dx$, $Var(X) = E[(x-\mu_x)^2]$. Moments $E(X^r)$ and mgf $m(t) = E(e^{tX})$
- Sampling distribution = distribution of a statistic
  - Statistic = Parameter + Bias + Chance error
  - Bias of sample mean and variance: $E(\bar{X}) = \mu,\ E(S^2) = \sigma^2$

  - Sampling distribution of mean and variance for a Normal population: $\bar{X}_n \sim N(\mu, \frac{\sigma^2}{n})$ and $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$

- Central Limit Theorem: for a random sample drawn from a population with mean μ, finite variance σ^2, then $\bar{X}_n \xrightarrow{d} N(\mu, \sigma/\sqrt{n})$. Typically n $\geq$ 25 is sufficient.
  - Application to Binomial: for X the sum of binomial successes: $X \sim N(np, np(1-p))$
- Student's t distribution: for a random sample from a Normal pop, $\frac{\bar{X}-\mu}{s/\sqrt{n}} \sim t_{n-1}$
- Statistical inference: Estimation
  - Point estimation
    * $MSE = E[(\hat{\theta} - \theta)^2] = Var(\hat{\theta}) + Bias(\hat{\theta})^2$
    * Unbaisedness, efficiency, consistency
  - Interval estimation:
    * For a Normal population, known variance: a $1-\alpha$ confidence interval for μ is $\bar{X} \pm Z_{\alpha/2}(\sigma/\sqrt{n})$
    * For a Normal population, unknown variance: a $1-\alpha$ confidence interval for μ is $\bar{X} \pm t_{\alpha/2,n-1}(s/\sqrt{n})$
    * Evaluate CI using length of the interval
    * If looking for a one-sided CI, one of the endpoints is $\pm\infty$ and the other is set using the α quantile of the relevant distribution, instead of the $\alpha/2$ quantile.

1.1 Statistical inference: Hypothesis testing

- Null vs. alternative hypotheses (in terms of population parameters)
- Critical regions and one vs. two-sided tests of the null hypothesis
- Types of errors: Type I = P(reject H0 | H0 correct) (controlled by α) and Type II = P(accept H0 | H0 false) (denoted by β)
- P-value = probability of observing a sample outcome at least as extreme as the one observed, under H0
- For a Normal population:
  - Known variance: test statistic $Z = \frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}$. Compare to standard normal quantiles.
  - Unknown variance: test statistic $t = \frac{\bar{X}-\mu_0}{s/\sqrt{n}}$. Compare to t_{n-1} distribution.
- Power of the test = P(reject H0 | H0 false) = $1-\beta$
  - For a given μ_0, α, n, one/two sided test, compute the power to show $\mu_A > \mu_0$ is true. When the variance is known.
  - For a one-sided test of $\mu_A > \mu_0$:

$$\beta = P[Z < Z_\alpha - \frac{|\mu_0 - \mu_A|}{\sigma/\sqrt{n}}] \rightarrow \text{power} = 1 - \beta$$

  - For a two sided test of $\mu_A \neq \mu_0$:

$$\beta \approx P[Z < Z_{\alpha/2} - \frac{|\mu_0 - \mu_A|}{\sigma/\sqrt{n}}] \rightarrow \text{power} = 1 - \beta$$

- Sample size: If $\Delta = |\mu_0 - \mu_A|$
  - For a one-sided test of $\mu_A > \mu_0$:
$$n = \frac{(Z_\alpha + Z_\beta)^2}{\Delta^2/\sigma^2}$$
  - For a two sided test of $\mu_A \neq \mu_0$:
$$n = \frac{(Z_{\alpha/2} + Z_\beta)^2}{\Delta^2/\sigma^2}$$
- Sample size and power curves. Plots power vs. parameter. Shows relationships between type of test, sample size, effect size, type I error, and standard deviation
- One sample inference: Proportions
  - Let $p = \bar{X} = \hat{\pi}, \sigma_p = \sqrt{\pi(1-\pi)/n} \hat{=} \sqrt{p(1-p)/n}$
  - A $1-\alpha$ CI for π is $p \pm Z_{\alpha/2}(\sigma_p)$. This approximation is fine if np and n(1-p) are both greater than 5 (by CLT)
  - A hypothesis test for π uses the test statistic $\frac{p-\pi_0}{\sigma_{p_0}}$, where σ_{p_0} is evaluated using π_0.
  - Can perform an exact hypothesis test using p-values calculated from the tails of the Binomial distribution under H0.
- One sample inference: Variance for Normal population
  - Since $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$, we can derive a 2-sided CI for σ^2 as
$$[\frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}}, \frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}}]$$
    Analogous result for a 1-sided interval, where the left endpoint is 0, using α.
  - For hypothesis testing, the test statistic is $\frac{(n-1)S^2}{\sigma_0^2}$, which we can compare to Chi-square quantiles.
- Two-sample inference: difference in means for independent groups $= \mu_1 - \mu_2$. Normal population.
  - For known variances,
$$\bar{X_1} - \bar{X_2} \sim N(\mu_1 - \mu_2, \sigma_1^2/n_1 + \sigma_2^2/n_2).$$
    Use this to construct CI the standard way using the standard normal distribution. Similarly for hypothesis testing, create test statistic using $H_0 : D_0 = \mu_1 - \mu_2$.
  - For unknown but equal variances, use pooled standard deviation $s_p = \sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}}$. Then the standard error becomes $s_p\sqrt{1/n_1 + 1/n_2}$. Use the t distribution with df $n_1 + n_2 - 2$. Similar for hypothesis testing.
  - For unknown and unequal variances:
    - Option 1: pretend population variances are known and equal sample variances (if $n_1, n_2 \geq 25$). Take this approach when hypothesis testing, and use degrees of freedom $\min(n_1 - 1, n_2 - 1)$.

  - Option 2: Use pooled variance and degrees of freedom $n_1 + n_2 - 2$. If $n_1 \approx n_2$ and $S^2_{max}/S^2_{min} \leq 3$

- Two-sample inference: difference in means for paired data. Normal population.
  - Consider individual differences $d_i = x_{1i} - x_{2i}$
  - Point estimate for $\mu_1 - \mu_2 = \bar{d} = \bar{X}_1 - \bar{X}_2$
  - For confidence intervals and hypothesis testing, use the standard error $s_d/\sqrt{n}$, where $s_d = \sqrt{\frac{\sum_i (d_i - \bar{d})^2}{n-1}}$. Then use $\bar{d}$ and the T distribution with n-1 df.
  - Power of a level α test (known variances):

$$\Phi(-Z_{1-\alpha/2} + \frac{\Delta}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}),$$

    where Δ is the meaningful difference of interest.
  - Sample size needed per group:

$$n = \frac{(\sigma_1^2 + \sigma_2^2)(Z_{1-\alpha/2} + Z_{1-\beta})^2}{\Delta^2}$$

- Multiple comparisons and Bonferroni approach. For m comparisons, make each comparison at the $\alpha = \alpha_0/m$ confidence level.

- Nonparametric methods:
  - Sign test
    - Analogue to single-sample t-test. Considering a hypothesis for the median θ.
    - $TS = \sum_{i=1}^n \mathbb{I}_{X_i > \theta_0} \sim Binom(n, 1/2)$. Then $Z = \frac{TS - n/2}{\sqrt{n/4}} \sim N(0,1)$.
    - Remove any observation that equals the median.
  - Wilcoxon signed rank test
    - Analogue to paired t-test. Considering pairs (X, Y). Assumes distributions for X and Y differ only by a shift in locaiton.
    - Rank absolute deviations $|d_j| = |x_j - y_j|$. Record the sign of the deviations and define $r_j = sgn(d_j) rank(|d_j|)$
    - $TS = r_j$ = (sum of ranks for positive) - (sum of ranks for negative)
    - Exclude observations with 0 deviation.
  - Wilcoxon rank sum test
    - Analogue to independent samples t-test. Assumes X, Y come from the same continuous population.
    - Rank all observations together. TS = sum of ranks of observations in smaller of the 2 samples.
  - Permutation tests

- Categorical data: Difference in proportions from 2 independent samples
  - CI: $(p_1 - p_2) \pm Z_{\alpha/2}(\sigma_{p_1 - p_2})$, where $\sigma_{p_1 - p_2} = \sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}}$. Can replace population proportions with sample proportions for estimator of variance.
  - Hypothesis test for $\pi_1 - \pi_2$: Use D_0, $\sigma_{p_1 - p_2} = \sqrt{\pi(1-\pi)(1/n_1 + 1/n_2)}$, $\hat{\pi} = \frac{\sum x_{1i} + \sum x_{2j}}{n_1 + n_2}$

- Categorical data: Difference in proportions from same sample (not independent)
  - Able to represent using a contingency table:

	B1	B0
A1	a	b
A0	c	d

    $p_A = \frac{a+b}{n}, p_B = \frac{a+c}{n}$ $p_A - p_B = \cdots = \frac{b}{n} - \frac{c}{n} = p'_A - p'_B$
  - CI for $\pi_A - \pi_B$: $(p_A - p_B) \pm Z_{\alpha/2}(\sigma_{p'_A - p'_B})$ where $\sigma_{p'_A - p'_B} = \sqrt{p'_A/n + p'_B/n - (p'_A - p'_B)^2/n}$
  - Hypothesis test of $\pi_A - \pi_B = 0$: Use $TS = \frac{p'_A - p'_B}{\sigma_{p'_A - p'_B, 0}}$, where $\sigma_{p'_A - p'_B, 0} = \sqrt{p'_A/n + p'_B/n}$
- Categorical data: Inference on contingency tables
  - Multinomial experiments and 1-way contingency tables. Test statistic $= \chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}, E_i = n\pi_{i0}$ summed over all k cells. Compare to Chi-square distribution with df (k-1). Null hypothesis = all cell probabilities equal hypothesized values: $\pi_i = \pi_{i0}\ \forall i$.
  - RxC tables (2-way tables). Test statistic $= \chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}, E_{ij} = \frac{n_{i.} n_{.j}}{n_{..}}$ summed over all cells. Compare to Chi-square distribution with df (r-1)(c-1). Null hypothesis = two variables are independent. Note the difference in degrees of freedom vs. a Chi-square goodness of fit test, which has cell proportions provided and degrees of freedom rc - 1.
  - 2x2 tables. Special case of above, with $\chi^2 = \frac{n_{..}(n_{11}n_{22} - n_{12}n_{21})^2}{n_{1.}n_{2.}n_{.1}n_{.2}}$, compared to df = (2-1)(2-1) = 1. This is equivalent to a Z test of independent samples. For small sample sizes, also have Fisher's exact test using hypergeometric distribution.
- Continuity corrections incorporated into normal approx of Binomial and Chi-square. Adding or subtracting 1/2.
- Poisson inference
  - Hypothesis test for single Poisson mean. Test stat is $TS = \sum X_i \sim Pois(n\mu)$. 1-sided p-value is then 2 * area in tail of Poisson dist in relevant extreme direction. Can also use Normal approximation from $TS \xrightarrow{d} N(n\mu_0, n\mu_0)$
  - Hypothesis test for difference in two Poisson means. Use normal approximation.
- Study design
  - Cohort (prospective studies)
  - Case-control (retrospective studies)
  - Cross-sectional (prevalence studies)
  - Randomized trials. Completely randomized design vs. randomized block design.
- Measures of risk
  - Relative risk (cohort, cross-sectional). p_1/p_2
  - Risk difference (cohort, cross-sectional). $p_1 - p_2$. Used as effect measure for power/sample size calcs.
  - Odds ratio (case-control). $\frac{p_1/(1-p_1)}{p_2(1-p_2)}$. Close to RR if diseases are rare. Used as effect measure for power/sample size calcs.

2 Biostat 200B

2.1 Simple linear regression

- Linear regression model of the form $E[Y|X_1,\ldots,_k] = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$
- Covariance $Cov(X,Y) = E[(X - E(X))(Y - E(Y)] = E(XY) - E(X)E(Y)$. Variance $Var(X) = Cov(X,X)$. Correlation $\rho_{XY} = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$ measures degree of *linear* relationship between X and Y.
- Sample covariance $\hat{Cov}(X,Y) = \frac{1}{n-1}\sum(x_i - \bar{x})(y_i - \bar{y})$. Sample correlation $r_{XY} = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2}\sqrt{\sum(y_i-\bar{y})^2}}$.
  If u and v are mean-centered vectors, then $r_{uv} = cos(\theta) = \frac{u'v}{|u||v|}$, where θ is the angle between the vectors.
- Simple linear regression. Model $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ with 1) ϵ_i has mean 0, constant variance σ^2 and 2) $Cov(\epsilon_i, \epsilon_j) = 0$. Note X is considered fixed.
  - Assumptions:
    - Linearity
    - Homoskedasticity
    - Uncorrelatedness
    - Then add $\epsilon_i \sim N(0, \sigma^2)$
  - Method of least squares: Minimize $\sum(Y_i - \beta_0 - \beta_1 X_i)^2$ w.r.t. the parameters.
    - Normal equations

$$1) \sum Y_i = n\beta_0 + \beta_1 \sum X_i$$

$$2) \sum X_i Y_i = \beta_0 \sum X_i + \beta_1 \sum X_i^2$$

    - Parameter estimates

$$\hat{\beta}_1 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2} = r_{XY}\frac{s_Y}{s_X} = \frac{Cov(X,Y)}{Var(X)}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \hat{X}$$

  - Coefficient is the *marginal* effect of X on Y
  - Residuals $e_i = Y_i - \hat{Y}_i = Y_i - (\beta_0 + \beta_1 X_i)$. Note that the true error (unlike the residuals) is unknown and unobservable.
  - Estimate conditional variance as:

$$\hat{\sigma^2} = \frac{1}{n-2}\sum(Y_i - \hat{Y}_i)^2 = MSE = \frac{1}{n-2}SSE$$

    MSE is an unbiased estimator of the conditional variance.
    - Estimator of standard deviation is $\sqrt{MSE}$
- Model diagnostics (SLR). Considering residuals, sometimes standardized residuals $e_i^* = \frac{e_i}{\sqrt{MSE}}$
  - Nonconstant error variance: Plot residual vs. fitted. Consequence: increased standard errors. Remedy: transform Y, WLS.
  - Nonlinearity of regression function: Plot residual vs. fitted. Consequence: misspecified mean function, coefficients are wrong and/or misleading. Remedy: transform X.

  - Residual normality: QQ/PP plot, univariate residual plot. Consequence: Validity of hypothesis tests + confidence intervals: Remedy: transform Y.
  - Correlated errors. Remedy: used mixed model.

- Inference for SLR
  - Gauss-Markov Theorem: under LR conditions, the least squares estimators are best (minimum variance) linear unbiased estimators
  - Note that $\hat{\beta}_1 = \sum c_i Y_i,\ c_i = \frac{x_i - \bar{x}}{\sum_{j=1}^{n}(x_j - \bar{x})^2}$ a linear combination of the outcome variable. So $\hat{\beta}_1$ is normal with mean and variance:

$$E(\hat{\beta}_1) = \beta_1$$

$$Var(\hat{\beta}_1) = \frac{\sigma^2}{\sum(x_i - \bar{x})^2} = \frac{\sigma^2}{(n-1)s_1^2}$$

    where s_1^2 is the sample variance of predictor X_1
  - Using MSE as estimate for the variance, we have

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\hat{\sigma^2}}{\sum(x_i - \bar{x})^2}}} \sim t_{n-2}$$

    Note $SE(\hat{\beta}_1)$ is a function of MSE (the estimated conditional variance), SST
  - Same for estimated intercept: $\frac{\hat{\beta}_0 - \beta_0}{SE(\hat{\beta}_0)} \sim t_{n-2},\ SE(\hat{\beta}_0) = \sqrt{\hat{\sigma^2}(\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2})}$. The standard error is a function of the conditional variance of the predictor and the sample mean.
  - Can use these to derive sampling dist of the conditional mean of interest, $\hat{Y_h} = \hat{\beta}_0 + \hat{\beta}_1 x_h$, and conduct inference.

- Mean centering: Can reduce SE of intercept if 0 is outside the range of the predictor values. Only changes the interpretation of the intercept, not the slope.

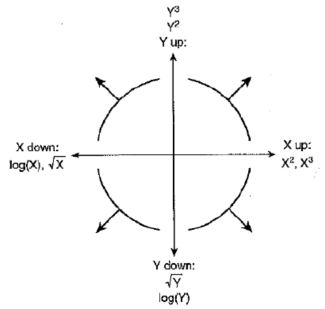
- Analysis of variance (SLR). Variance decomposes as $Y_i - \bar{Y} = (Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y})$
  - Sum of squares: SST = SSE + SSR

$$\sum(Y_i - \bar{Y})^2 = \sum(Y_i - \hat{Y}_i)^2 + \sum(\hat{Y}_i - \bar{Y})^2$$

    * SST = total variation in Y
    * SSE = unexplained variation in Y
    * SSR = variation in Y explained by linear association with X
  - Null distributions and degrees of freedom (if no regression relation $\iff \beta_1 = 0$)
    * $\frac{SSTO}{\sigma^2} \sim \chi^2_{n-1} \implies$ SST has n-1 df.
    * $\frac{SSE}{\sigma^2} \sim \chi^2_{n-2} \implies$ SSE has n-2 df (note: why is this true whether or not $\beta_1 = 0$?)
    * $\frac{SSR}{\sigma^2} \sim \chi^2_1 \implies$ SSR has 1 df
  - Mean squares are SS divided by df
  - Hypothesis test of $H_0 : \beta_1 = 0$ has test statistic $F = \frac{MSR}{MSE} = \frac{SSR/df_{Reg}}{SSE/df_E} \sim F_{1,n-2}$ (overall/omnibus F test). Level α test rejects null hypothesis if $F > F_{1-\alpha;1,n-2}$. Note for SLR this is equivalent to a t test: $F = t^2$

  - $R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$. Measures proportion of variation in Y explained by linear association with X. For SLR, sample correlation $r = \pm\sqrt{R^2}$

- Power transformations: For nonlinear relationship between Y and X
  - $x \mapsto x^p$. $p > 1$: larger values spread more than smaller values. $p < 1$: larger values spread less than smaller values. For $p < 0$, also multiply by -1 to preserve pos/neg relationship
  - Circle of transformations: plot X vs. Y, go up or down ladder of powers for X/Y based on direction of bulge



  - Can only use power transformations for monotone and/or simple relationships (without min/max in range of values)
  - Transform Y to deal with heteroskedasticity. Transform X for better interpretation.

- Loess curve. Scatterplot smoother, estimate $E[Y|x = x_i]$ only using values in a neighborhood of x_i, weighting closer observations more than farhter observations. Increase span = smoother curve.

2.2 Multiple linear regression

- Multiple linear regression. Model $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \epsilon_i,\ \epsilon_i \sim iid\ N(0, \sigma^2)$
  - Estimated using least squares
  - Coefficients are *partial* effects of X_i on Y, holding all other covariates constant
  - Residuals are vertical distances between Y_i and the fitted values on the p+1 dimension hyperplane in $\mathbb{R}^{p+2}$. (p+1 = number of parameters including intercept, p + 2 includes dependent variable)
  - Estimate $\hat{\sigma^2} = MSE = \frac{SSE}{df_{SSE}} = \frac{1}{n-p-1}\sum(Y_i - \hat{Y_i})^2$
  - Anova table:

Source	SS	df	MS	F
Regression	$SSReg$	k	$MSReg = SSReg/k$	$MSReg/MSE$
Error	SSE	$n-(k+1)$	$MSE = SSE/(n-k-1)$	
Total	$SSTO$	$n-1$		

  - Interpretation of partial regression coefficient: estimate partial relationship after linear relationship between other predictors are accounted for. Example for 2-predictor model with intercept: 1) Regress Y on X_2, get residuals $e_{i,y|X_2}$. 2) Regress X_1 on X_2, get residuals $e_{i,X_1|X_2}$. 3) $\hat{\beta}_1$ is the slope coefficient from regression of $e_{i,y|X_2}$ on $e_{i,X_1|X_2}$
  - Partial correlation coefficient: $Corr(e_{i,y|X_2,\ldots,X_p}, e_{i,X_1|X_2,\ldots,X_p})$ = correlation between Y and X_1 after adjusting for other covariates
  - Can present results using adjusted means, the outcome evaluated at the sample averages for each covariate

- Dummy variables: use c-1 binary variables to encode means for c categories
- Inference for MLR
  - Single coefficient tests:

$$\frac{\hat{\beta}_j - \beta_j^*}{\sqrt{\hat{Var}(\hat{\beta}_j)}} \sim t_{n-(k+1)}$$

    Confidence interval:

$$\hat{\beta}_j \pm t_{1-\alpha/2,n-(k+1)}\sqrt{\hat{Var}(\hat{\beta}_j)}$$

    Variance of $\hat{\beta}_j$ in linear algebra section below.
  - Simultaneous/joint tests for multiple coefficients: Consider whether increase in SSR (decrease in SSE) from adding predictors explains a lot of variation in Y, beyond the variation from predictors already included.
    * Testing null hypothesis that additional coefficients equal 0.
    * $SSE(X_1,\ldots X_{m-1}) - SSE(X_1,\ldots X_{m-1},X_m,\ldots X_j) = SSR(X_1,\ldots X_{m-1},X_m,\ldots X_j) - SSR(X_1,\ldots X_{m-1})$
    * Partial F test for full (F) vs. reduced (R) model:

$$F^* = \frac{\frac{SSE(R)-SSE(F)}{df_{SSE}(R)-df_{SSE}(F)}}{\frac{SSE(F)}{df_{SSE}(F)}} = \frac{\frac{SSR(F)-SSR(R)}{df_{SSR}(F)-df_{SSR}(R)}}{\frac{SSE(F)}{df_{SSE}(F)}} \sim F_{df_{SSE}(R)-df_{SSE}(F),df_{SSE}(F)}$$

    * Difference in degrees of freedom in numerator = number of parameters being tested equal 0
    * Denominator is always MSE(F)
    * F test for single coefficient is equivalent to t-test of 0
  - Overall F test: tests whether all slope coefficients equal 0 (not the intercept). Reduced model: $\hat{Y}_i = \bar{Y}$ (SSE(R) = SST).

$$F^* = \frac{\frac{SST-SSE(F)}{(n-1)-(n-k-1)}}{\frac{SSE(F)}{n-k-1}} = \frac{\frac{SSR(F)}{k}}{\frac{SSE(F)}{n-k-1}} = \frac{MSR}{MSE}$$

- Polynomial regression. Model $Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_{ij}^j + \epsilon_i$.
  - Quadratic polynomial: use when min/max is included in range of values.
  - Interpretation of coefficient for quadratic: $dY/dX = \beta_1 + 2\beta_2 X \implies \beta_1$ is slope of tangent line when X = 0
  - Power transformations are ok only when data values are positive with monotone relationship. Prefer a quadratic over power transformation in general.

- Log transformations. Log moves down the ladder of powers. Often used for positive skew. Reflects multiplicative changes.
  - $\log_{10} X \implies$ unit increase is multiplying X by 10. $\log_2 X \implies$ unit increase is doubling X
  - Natural log transform of X: 1% increase in X associated with increase in mean Y of $0.01\hat{\beta}_1$
  - Natural log transform of Y: 1 unit increase in X associated with change in mean Y by a *factor of* $e^{\hat{\beta}_1}$
  - Log X and log Y: 1 percent increase in X associated with β_1% increase in Y
- Interactions: Effect of one predictor X on outcome Y depends on value of Z. Interaction effects = "moderation", "effect modifiers" = conditional effects.
  - Note correlation and interaction are distinct.
  - Test significance for higher order terms first, keep model hierarchically well-formulated
  - Mean-centering can help with interpretation and decrease standard errors.
  - Continuous X + categorical Z: without interaction, effect of X is same for all levels of Z (different intercepts, common slope). With interaction, effect of X varies by level of Z (different intercepts, different slopes). Interpretation of lower-order terms are now partial effects conditional on value of other variable = 0.
  - Continuous X and Z: Without interaction, regression surface is a plane. With interaction, now surface is not a plane. Interaction coefficient is change in slope of Z associated with one-unit increase in X (and vice-versa)
- Continuous decay model. Example of nonlinear regression:

$$\frac{dY}{dx} = -kY \implies Y = Y_0 e^{-kx} \implies \ln Y = \ln Y_0 - kx \implies Y = \beta_0 + \beta_1 x$$

  Can solve for half-life $t_{1/2} = \ln 2/k$
- Multicollinearity. If predictors are uncorrelated, SLR coefs = MLR coefs.
  - Compute $R_j^2 = R^2$ obtained from regressing X_j on other predictors. Then:

$$Var(\hat{\beta}_j) = \frac{1}{(1-R_j^2)} \frac{\sigma^2}{\sum_{i=1}^n (X_{ij} - \bar{X}_j)^2} = \frac{1}{(1-R_j^2)} \frac{\sigma^2}{(n-1)s_j^2}$$

    where s_j^2 is sample variance of x_j
  - Variance inflation factor for predictor j: $VIF_j = \frac{1}{1-R_j^2}$. Use $\sqrt{VIF_j}$ for impact of collinearity on $SE(\hat{\beta}_j)$. Tolerance $TOL_j = 1/VIF_j$. Rule of thumb for VIF is > 10 is problematic.
- Linear algebra formulation for MLR
  - Model: $Y = X\beta + \varepsilon$, $E(\varepsilon) = 0$, $Var(\varepsilon) = \sigma^2 I_n$, where $Y \in \mathbb{R}^n, X \in \mathbb{R}^{n \times p}, \beta \in \mathbb{R}^p, \varepsilon \in \mathbb{R}^n$. With Normality of errors, we have $Y \sim N_n(X\beta, \sigma^2 I)$
  - Least squares and the normal equation: $Q(\beta) = (Y - X\beta)^T (Y - X\beta)$. $\frac{\partial Q}{\partial \beta} = 0 \implies X^T Y = X^T X \beta$. Left multiply by the inverse of the gram matrix to get $\hat{\beta} = (X^T X)^{-1} X^T Y$

- Fitted values and the hat matrix: $\hat{Y} = X\hat{\beta} = X(X^TX)^{-1}X^TY = HY$ for $H = X(X^TX)^{-1}X^T$. The hat matrix is symmetric and idempotent: $H^T = H, HH = H$. Also $I - H$ symmetric and idempotent.
- Variance of Y and residuals: $Var(\hat{Y}) = H\ Var(Y)H^T = \cdots = \sigma^2 H$. Residuals $e = (I - H)Y \in \mathbb{R}^n$. $Var(e) = \sigma^2(I - H)$.
- Gram matrix X^TX. Recall row rank = col rank, and rank(A) = rank(A^TA). Then $1 \leq \text{rank}(X^TX) = \text{rank}(X) \leq p$, assuming $n > p$. Inverse of the gram matrix exists only if it's full rank $\iff$ X has p linearly independent columns. If $n < p$, can use generalized inverse A^- satisfying $AA^-A = A$
- ANOVA results. All quadratic forms
  - $SSE = (Y - HY)^T(Y - HY) = Y^T(I - H)Y$. $Rank(I - H) = n - p$
  - $SST = (Y - \frac{1}{n}JY)^T(Y - \frac{1}{n}JY) = Y^T(I - \frac{1}{n}J)Y$, for $J = \mathbb{1}_n\mathbb{1}_n^T$. $Rank(I - \frac{1}{n}J) = n - 1$
  - $SSR = (HY - \frac{1}{n}JY)^T(HY - \frac{1}{n}JY) = Y^T(H - \frac{1}{n}J)Y$. $Rank(H - \frac{1}{n}J) = p - 1$
  - Theorem: if $Y \sim N(0, \sigma^2 I)$ and M symmetric idempotent matrix with rank m, then $\frac{1}{\sigma^2}Y^TMT \sim \chi^2_m$
  - Overall F test:
$$F = \frac{\frac{1}{\sigma^2}\frac{SSR}{df_{Reg}}}{\frac{1}{\sigma^2}\frac{SSE}{df_{Err}}} = \frac{\frac{1}{\sigma^2}\frac{Y^T(H - \frac{1}{n}J)Y}{p-1}}{\frac{1}{\sigma^2}\frac{Y^T(I-H)Y}{n-p}} \sim F_{p-1,n-p}$$
- Inference
  - Coefficients are multivariate normal: $\hat{\beta} \sim N_p(\beta, \sigma^2(X^TX)^{-1})$. So individual coefficients are univariate normal, and inference for individual coefficients use t distribution:
$$\frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}\sqrt{v_{jj}}} \sim t_{n-p}$$
    where v_{jj} is the jth diagonal element of $(X^TX)^{-1}$ (same formula as above section on MLR inference)
  - For multiple coefficients, if $\hat{\beta}_1$ are the coefficients being tested, and V_{11} is the square submatrix of the covariance matrix for these coefficients:
$$F = \frac{\frac{(\hat{\beta_1} - \beta_1^0)^T V_{11}^{-1}(\hat{\beta_1} - \beta_1^0)}{q}}{\frac{SSE(F)}{n-p}}$$
    For the omnibus F test:
$$F = \frac{\frac{\hat{\beta}^T(X^TX)^{-1}\hat{\beta_1})}{p-1}}{\frac{SSE(F)}{n-p}}$$
- Evaluating mean response at covariate values: $\hat{Y_h} = x_h^T\hat{\beta}$, with $\hat{Var}(\hat{Y_h}) = \hat{\sigma^2}x_h^T(X^TX)^{-1}x_h$. Can use for inference on adjusted means.

- Multivariate Normal distribution
- Outliers & residuals
  - X-outliers: measured by Leverage = diagonal entries of hat matrix
  - Conditional Y-outliers/regression outliers: measured by Residual
  - Raw residuals: $e_i = y_i - \hat{y_i}$

  - Standardized residual = raw residual standardized by its SD: $r_i = \frac{e_i}{\hat{\sigma}\sqrt{1-h_i}}$. Have variance 1, in units of SD.
  - Standardized/studentized deleted residuals = using fitted value from regression coefficients only using n-1 observations: $t_i = e_i[\frac{n-p-1}{SSE(1-h_i)-e_i^2}]^{1/2}$
  - Leverage for observation i: $h_i = h_{ii} = x_i^T(X^TX)^{-1}x_i$
    * For SLR: $h_i = \frac{1}{n} + \frac{(x_i-\bar{x})^2}{\sum(x_i-\bar{x})^2}$. Measure of distance from $\bar{x}$
    * For MLR: $h_i = \frac{1}{n} + x_i^{*T}(X^{*T}X^*)^{-1}x_i^*$, where x_i^* is the ith observation mean-centered
    * Has properties: bounded $\frac{1}{n} \leq h_i \leq 1$, sum $\sum h_i = rank(X)$, average $\bar{h} = \frac{p}{n}$
    * High if larger than 2 or 3 times $\bar{h}$
  - Influence: function of X and Y outlyingness
    * Cook's D: $D_i = \frac{\sum(\hat{y}_i - \hat{y}_{i(-i)})^2}{p\hat{\sigma^2}} = \frac{e_i^2 h_i}{p\hat{\sigma}(1-h_i)^2}$
    * Cook's D is high if $D_i > 4/(n-p)$

- Partial relationship plots. Evaluating linearity in partial relationship with Y for each predictor in MLR
  - Partial regression/added-variable plots: Plot residuals from partial regressions: $e_{y|x_2,\ldots,x_{p-1}}$ vs $e_{x_1|x_2,\ldots,x_{p-1}}$
  - Component-plus-residual/partial residual plot: Plot partial residuals for jth predictor: $e_i^{(j)} = \hat{\beta}_j x_{ij} + e_i$. Adding back linear (or quadratic, etc) relationship for jth predictor to residual.

- Variable selection: Inference
  - Bivariate p-value threshold: choose predictors with SLR p-values lower than some threshold
  - Adjusted R^2 : $R^2_{adj} = 1 - \frac{SSE/(n-p)}{SST/(n-1)} = 1 - \frac{MSE}{s_y^2}$. Dividing by degrees of freedom to account for number of parameters. Prefer higher value (note regular R2 will always increase with more predictors). Equivalent to using MSE. Can lead to too many predictors.
  - Mallow's Cp: $Cp = \frac{SSE(R)}{MSE(F)} - (n-2p)$, where the full model is using the entire pool of predictors. Estimates standardized sum of squared differences of fitted vs. true (unobserved) conditional means. Prefer lower Cp.
  - Akaike Information Criterion (AIC): $AIC = -2\log L(\hat{\theta}) + 2p = n\log(\hat{\sigma^2}) + 2p$ for MLR. Prefer lower AIC.
  - Bayesian Information Criterion (BIC): $BIC = -2\log L(\hat{\theta}) + \log(n)p = n\log(\hat{\sigma^2}) + \log(n)p$. Prefer lower BIC.
  - Care about differences in AIC/BIC relative to other models, not magnitude.

- Variable selection: Prediction
  - Prediction MSE: $\frac{1}{n}\sum(y_i - \hat{y}_i)^2$. Differs from Anova MSE in denominator.
  - Train/test set, validation set $\implies$ train/test/validation MSE
  - Cross-validation (LOO, k-fold). Test MSE estimated as average of k fold-MSEs.
  - Best subset selection: fit all possible models
  - Stepwise selection: forward, backward, forward-backward. Add or subtract based on improvements to a criterion or statistical significance. Can lead to overfitting, not guaranteed to find best model.

- Splines
  - Basis functions: family of functions used to model X through a linear combination. Polynomial regression for example.
  - Step functions: piecewise constant between k knots. Model: $E(y_i) = \beta_0 + \beta_1 C_1(x_i) + \cdots + \beta_k C_k(x_i)$, indicator functions as basis functions.
  - Piecewise linear regression: separate regression function between each knot, can impose continuity
  - Regression spline: piecewise polynomial between each knot which is continuous and smooth at the knots. Cubic spline most popular.
    - Model with two knots: $E(y_i) = \beta_0 + \beta_{11}x_{i1} + \beta_{12}x_{i1}^2 + \beta_{13}x_{i3}^3 + \beta_{23}x_{i2}^3 + \beta_{33}x_{i3}$
    - With k knots, have k + 1 bins, with k + 4 parameters: 1) one intercept, 2) 3 for linear, quadratic, cubic within first bin, 3) k remaining cubic terms in remaining k bins
    - Natural/restricted cubic splines: constrain regression function to be linear outside of boundary knots.
    - Advantage: still linear in parameter and can use least squares, can incorporate directly into a regression model. Disadvantage: not easily interpretable and need plots to visualize
  - Smoothing spline: find function $g(x)$ to minimize loss + penalty = $\sum(y_i - g(x_i))^2 + \lambda \int g''(t)^2 dt$, with tuning parameter λ. Note: using L2 loss, and second derivative in the penalty is a measure of roughness of function. Values of λ correspond to effective degrees of freedom.
- Regression to the mean. In context of repeated measures, if the first measure is extreme, the second measure will tend to be closer to the population mean. Since most probability mass is near the mean, using an cutoff for eligibility criteria will select some individuals who are naturally extreme as a result of natural variation.
  - If X is first measurement and Y is second measurement, from regressing Y on X:

    $$\frac{\hat{y}_i - \bar{y}}{s_y} = r\frac{x_i - \bar{x}}{s_x}$$

    Y is closer to the mean than X is, since $|r| < 1$
  - Percent regression to the mean = $100(1 - |r|)$
  - For a sample selected using a cutoff, RTM effect = $(1 - |r|)|\mu_{pop} - \mu_{sample}|$
- Nonconstant error variance
  - Regression assumption $Var(\varepsilon) = \sigma^2 I$ guarantees BLUE from Gauss-Markov Theorem.
  - If heteroskedasticity, have $Var(\varepsilon) = \Sigma = diag(\sigma_1^2, \ldots, \sigma_n^2)$. OLS estimate is still unbiased, but $Var(\hat{\beta}) = (X^TX)^{-1}X^T\Sigma X(X^TX)^{-1} \neq \hat{\sigma}^2(X^TX)^{-1}$. Results in wrong standard errors, usually overestimated.
  - A problem when ratio of largest to smallest error variances is 4 or more
  - Transformations of Y
  - Weighted least squares (WLS).
    - Divide each observation by the SD of its error term, so $Var(\varepsilon_i^{(w)}) = 1$.
    - Define $W = diag(1/\sigma_1^2, \ldots, 1/\sigma_n^2)$ s.t. $W^{-1} = \Sigma$, and $W^{1/2} = diag(1/\sigma_1, \ldots, 1/\sigma_n)$

  - Model: $W^{1/2}Y = W^{1/2}X\beta + W^{1/2}\varepsilon$
  - Estimator $\hat{\beta}^{(w)} = (X^TWX)^{-1}X^TWY$, with $E(\hat{\beta}^{(w)}) = \beta, Var(\hat{\beta}^{(w)}) = (X^TWX)^{-1}$
  - In practice, set weights $w_i = 1/e_i^2$, where e_i is residual from OLS. This is a strong assumption.
- Sandwich estimator of Σ. Use estimate of the covariance matrix $\Sigma = \hat{\Sigma} = diag(e_1^2, \ldots, e_n^2)$ and plug in to estimator for $Var(\hat{\beta})$. The goal is to be robust to covariance structure misspecification, not to reduce standard errors.

2.3 ANOVA

2.3.1 One-way ANOVA

Comparing group means when classified by one factor. Equivalent to regression with dummy variables.

- Model: $y_{ij} = \mu_i + \varepsilon_{ij},\ \varepsilon_{ij} \sim N(0, \sigma^2)$, with levels $i = 1, \ldots, a$ and observations within each group $j = 1, \ldots, N_i$. Parameters are group means and common variance. Estimate them with sample group means and pooled sample variance s_p^2:

$$\hat{\mu_i} = \frac{1}{N_i}\sum_j y_{ij}$$

$$s_i^2 = \frac{1}{N_i - 1}\sum_j (y_{ij} - \bar{y}_i)^2$$

$$s_p^2 = \frac{\sum_i (N_i - 1)s_i^2}{\sum_i N_i - 1} = \frac{1}{n-a}[\sum_{i=1}^{a}\sum_{j=1}^{N_i}(y_{ij} - \bar{y}_i)^2] = MSE$$

- Variance decomposition: $y_{ij} - \bar{y} = (y_{ij} - \bar{y}_i) + (\bar{y}_i - \bar{y}_{..})$.
  - Sum of squares: SST = SSE + SSGroup

$$\sum_{i=1}^{a}\sum_{j=1}^{N_i}(y_{ij} - \bar{y})^2 = \sum_{i=1}^{a}\sum_{j=1}^{N_i}(y_{ij} - \bar{y}_i)^2 + \sum_{i=1}^{a}\sum_{j=1}^{N_i}(\bar{y}_i - \bar{y}_{..})^2$$

  - SST: n-1 df, SSE: n-a df, SSGroup: a-1 df
  - $F = \frac{MSGroup}{MSE} = \frac{SSGroup/(a-1)}{SSE/(n-a)}$
- Inferences about means
  - Test equality of all means: $H_0 : \mu_1 = \cdots = \mu_a = \mu$. Equivalent to omnibus F test, use above F statistic.
  - Inference for single group mean. Perform t-test: $Var(\bar{y}_i) = \frac{\sigma^2}{N_i} \implies \hat{Var}(\bar{y}_i) = \frac{s_p^2}{N_i} \implies \frac{\bar{y}_i}{\sqrt{\frac{s_p^2}{N_i}}} \sim t_{n-a}$
  - Testing contrasts (linear combination of means with coefficients summing to 0): $\sum_i \lambda_i\mu_i$, s.t. $\sum_i \lambda_i = 0$. Estimate is $\sum_i \lambda_i\bar{y}_i$, which is unbiased with variance $\sigma^2\sum_i \frac{\lambda_i^2}{N_i}$. Test that the contrast (the "parameter") equals 0:

$$\frac{\sum_i \lambda_i\bar{y}_i}{\sqrt{MSE\sum_i \frac{\lambda_i^2}{N_i}}} \sim t_{n-a}$$

    Reject H_0 if TS $> t_{1-\alpha,n-a}$. Equivalent to squaring the test statistic and expressing as sum of squares of the parameter, and using the $F_{1,n-a}$ distribution.

- Checking model assumptions
  - Common variance: compare sample variances for each level
  - Normality: QQ plot of residuals $\hat{\varepsilon} = y_{ij} = \bar{y}_i$
- Lognormal distribution: If transformed observations $Y_1 = \log(X_1), \ldots, Y_n = \log(X_n) \sim iidN(\mu, \sigma^2)$, the original observations are lognormal: $X_1, \ldots, X_n \sim iid \log N(\mu, \sigma^2)$. Parameters are the mean/variance after log-transformation.
  - Geometric mean = median: $\gamma = exp(\mu)$

    $$\bar{Y} = \frac{1}{n}\sum_i log(X_i) \implies \exp(\bar{Y}) = (\prod_i X_i)^{1/n}$$

    Exponentiating log transformed data gives median/geometric mean in original scale. Difference in means on log-transformed scale is ratio of medians/geometric means on original scale: $\exp(\mu_1 - \mu_2) = \gamma_1/\gamma_2$
  - Coefficient of variation (CV = SD/mean): $\sqrt{\exp(\sigma^2) - 1}$
- Unequal variances across levels
  - Test $H_0 : \sigma_1^2 = \cdots = \sigma_a^2$. Levene's test: Conduct one-way anova on deviations $d_{ij} = |Y_{ij} - \bar{Y}_i|$. Brown-Forsythe test uses medians instead.
  - If ratio of largest to smallest variances < 3, it's ok. Only care about large departures from variance homogeneity, and about very low significant levels for the above tests.
  - Can use transformations like log
  - Welch's ANOVA test/Satterthwaite approximation. Used if normality assumption is met. To test contrasts, use test statistic:

    $$\frac{\sum c_i \bar{Y}_i}{\sqrt{\sum c_i^2 \frac{\hat{\sigma}_i^2}{n_i}}} \sim t_{df}$$

    with $df = \frac{(\sum c_i^2 \frac{\hat{\sigma}_i^2}{n_i})^2}{\sum(\frac{\sum c_i^2 \frac{\hat{\sigma}_i^2}{n_i}}{n_i - 1})}$
  - Nonparametric one-way anova. Kruskal-Wallis test: using ranks.
- Factor effects model
  - Parametrization: $y_{ij} = \mu + \tau_i + \varepsilon_{ij},\ \varepsilon_{ij} \sim N(0, \sigma^2)$, with levels $i = 1, \ldots, a$ and observations within each group $j = 1, \ldots, N_i$. Cell means are $\mu_i = \mu + \tau_i$. Now have a + 1 parameters for a levels (plus variance).
  - Zero-sum constraint: $\sum_i \tau_i = 0$. Enforces that $\mu_a = \mu - \sum_{i=1}^{a-1} \tau_i$ and reduces number of parameters by 1. Ensures model is not overparametrized (design matrix not full rank). Different generalized inverses correspond to different constraints; the zero-sum constraint is using a particular generalized inverse $(X^T X)^-$
  - Design matrix and model for 3 groups, 2 obs per group: $Y = X\beta + \varepsilon$

    $$Y = \begin{bmatrix} Y_{11} \\ Y_{12} \\ Y_{21} \\ Y_{22} \\ Y_{31} \\ Y_{32} \end{bmatrix},\ X = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & -1 & -1 \\ 1 & -1 & -1 \end{bmatrix},\ \beta = \begin{bmatrix} \mu \\ \tau_1 \\ \tau_2 \end{bmatrix},\ \epsilon = \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \varepsilon_{31} \\ \varepsilon_{32} \end{bmatrix}$$

  - Interpretation. Grand mean = mean of means: $\mu = \frac{1}{a}\sum \mu_i$. Factor/treatment effects = difference from grand mean: $\tau_i = \mu_i - \mu$
  - Orthogonal contrasts: for two contrasts with coefficients c and d, the contrasts c and d are orthogonal if $c^T d = 0$. Estimates from orthogonal contrasts are statistically independent. For a groups, have a-1 orthogonal contrasts. Can also use orthogonal polynomial contrasts to test for trends among the means or to encode a polynomial regression.

- Multiple comparisons. M tests, want to control experimentwise/familywise error rate (EER). For one test, this is this comparisonwise error rate (CER)
  - Bonferroni: Use CER of α/m for m tests
  - Sidak: Use CER of $1 - (1-\alpha)^{1/m}$
  - For one-way ANOVA, we have $m = \binom{a}{2} = \frac{a(a-1)}{2}$ possible comparisons between all mean pairs, and $m = a - 1$ possible comparisons between a control and all other means. We can change the confidence coefficient (critical value) to compare mean differences used for null hypothesis test rejection.
  - Fisher's least significant difference.

    $$LSD = t_{1-\alpha/2,N-a}\sqrt{MSE(1/n + 1/n)}$$

    Does not control EER. Protected LSD: 1) first conduct omnibus F test, 2) if fail to reject, stop, 3) if reject null, use confidence coefficient $t_{N-a,1-\alpha/2}$. Protected version also does not control EER.
  - Bonferroni. Use critical value $t_{N-a,1-\alpha/(2m)}$:

    $$BSD = t_{N-a,1-\alpha/(2m)}\sqrt{MSE(1/n + 1/n)}$$

  - Tukey-Kramer. For making all pairwise comparisons. Based on studentized range distribution: $\frac{\bar{Y}_{max} - \bar{Y}_{min}}{s_p/\sqrt{n}} \sim$ s.r.d.$(a, df_{s_p} = N - a)$, for a samples of size n. Controls EER at α:

    $$HSD = q_{1-\alpha,a,N-a}\sqrt{MSE(1/n)}$$

  - Dunnett. For comparing each group i to a single control group 0. Comparing below test statistic to a critical value from Dunnett test table.

    $$t_{i0} = \frac{\bar{y}_i - \bar{y}_0}{s_p\sqrt{1/n_i + 1/n_0}}$$

  - Scheffe. Provides simultaneous CIs for all possible linear contrasts among means. Contrast CIs are:

    $$\sum_i c_i \bar{Y}_{i\cdot} \pm \sqrt{aF_{1-\alpha,a,N-a}}SE(\sum_i c_i \bar{i}\cdot$$

  - Simultaneous confidence level = probability all m CIs include the true parameter values

2.3.2 Two-way ANOVA

Cross-classifying by two factors.

- Blocking factor = observational variable that is a source of variability. Want to control for it and not estimate it.
- Model: $y_{ijk} = \mu_{ij} + \varepsilon_{ijk},\ \varepsilon_{ijk} \sim N(0, \sigma^2)$, with levels $i = 1, \ldots, a,\ j = 1, \ldots, b$. Relationships among means can be additive or with interactions

		B			
		$j=1$	$j=2$	$j=3$	Row means
A	$i=1$	$\mu_{11}=\mu+\alpha_1+\beta_1$	$\mu_{12}=\mu+\alpha_1+\beta_2$	$\mu_{13}=\mu+\alpha_1+\beta_3$	$\mu_{1\cdot}=\mu+\alpha_1$
	$i=2$	$\mu_{21}=\mu+\alpha_2+\beta_1$	$\mu_{22}=\mu+\alpha_2+\beta_2$	$\mu_{23}=\mu+\alpha_2+\beta_3$	$\mu_{2\cdot}=\mu+\alpha_2$
	Col means	$\mu_{\cdot 1}=\mu+\beta_1$	$\mu_{\cdot 2}=\mu+\beta_2$	$\mu_{\cdot 3}=\mu+\beta_3$	μ

- Additive effects. Sum of grand mean with main effects: $\mu_{ij}=\mu+\alpha_i+\beta_j$ with zero-sum constraints $\sum_i \alpha_i=\sum_j \beta_j=0$.
  - Estimates. Least squares (equal cell sizes): $\sum_i\sum_j\sum_k[(y_{ijk}-(\mu+\alpha_i+\beta_j)]^2 \implies$

$$\hat{\mu}=\bar{Y}_{\cdots}, \hat{\alpha}_i=\bar{Y}_{i\cdot\cdot}-\bar{Y}_{\cdots}, \hat{\beta}_j=\bar{Y}_{\cdot j\cdot}-\bar{Y}_{\cdots} \implies$$

$$\hat{\mu}_{ij}=\bar{Y}_{i\cdot\cdot}+\bar{Y}_{\cdot j\cdot}-\bar{Y}_{\cdots}$$

    Error variance: $MSE=\hat{\sigma}^2=\frac{1}{N-a-b+1}\sum_i\sum_j\sum_k[(y_{ijk}-(\hat{\mu}+\hat{\alpha}_i+\hat{\beta}_j)]^2$
  - Variance decomposition (equal cell sizes): $Y_{ijk}-\bar{Y}_{\cdots}=(Y_{ijk}-\bar{Y}_{ij\cdot})+(\bar{Y}_{ij\cdot}-\bar{Y}_{\cdots})$. SST = SSE + SSA + SSB:

$$\sum_i\sum_j\sum_k(Y_{ijk}-\bar{Y}_{\cdots})^2=\sum_i\sum_j\sum_k(Y_{ijk}-\bar{Y}_{ij\cdot})^2+\sum_i\sum_j\sum_k(Y_{i\cdot\cdot}-\bar{Y}_{\cdots})^2+\sum_i\sum_j\sum_k(Y_{\cdot j\cdot}-\bar{Y}_{\cdots})^2$$

  - Inference. Test that all levels of A have same effect on response: $H_0^A:\alpha_1=\cdots=\alpha_a=0$ using F test: $F_A=\frac{MSA}{MSE}\sim F_{a-1,N-a-b+1}$ under H_0^A. Same for factor B.

Source of variation	df	Sum of squares	Mean square	F
Factor A	$a-1$	$SSA=\sum_i\sum_j\sum_k(\bar{Y}_{i\cdot\cdot}-\bar{Y}_{\cdots})^2$	$MSA=\frac{SSA}{a-1}$	$F_A=\frac{MSA}{MSE}$
Factor B	$b-1$	$SSB=\sum_i\sum_j\sum_k(\bar{Y}_{\cdot j\cdot}-\bar{Y}_{\cdots})^2$	$MSB=\frac{SSB}{b-1}$	$F_B=\frac{MSB}{MSE}$
Error	$N-a-b+1$	$SSE=\sum_i\sum_j\sum_k(Y_{ijk}-\bar{Y}_{ij\cdot})^2$	$MSE=\frac{SSE}{N-a-b+1}$	
Total	$N-1$	$SSTOT=\sum_i\sum_j\sum_k(Y_{ijk}-\bar{Y}_{\cdots})^2$		

- Interaction effects. Sum of grand mean, main effects, and interaction effect: $\mu_{ij}=\mu+\alpha_i+\beta_j+(\alpha\beta)_{ij}$ with zero-sum constraints $\sum_i\alpha_i=\sum_j\beta_j=\sum_i(\alpha\beta)_{ij}=\sum_j(\alpha\beta)_{ij}=0$.
  - Estimates. Least squares (equal cell sizes): $\sum_i\sum_j\sum_k[(y_{ijk}-(\mu+\alpha_i+\beta_j+(\alpha\beta)_{ij})]^2 \implies$

$$\hat{\mu}=\bar{Y}_{\cdots}, \hat{\alpha}_i=\bar{Y}_{i\cdot\cdot}-\bar{Y}_{\cdots}, \hat{\beta}_j=\bar{Y}_{\cdot j\cdot}-\bar{Y}_{\cdots}, \widehat{(\alpha\beta)}_{ij}=\bar{Y}_{ij\cdot}-(\bar{Y}_{i\cdot\cdot}+\bar{Y}_{\cdot j\cdot}-\bar{Y}_{\cdots})$$

    Error variance: $MSE=\hat{\sigma}^2=\frac{1}{N-ab}\sum_i\sum_j\sum_k[(y_{ijk}-(\hat{\mu}+\hat{\alpha}_i+\hat{\beta}_j+\widehat{(\alpha\beta)}_{ij})]^2$
  - Variance decomposition (equal cell sizes): $Y_{ijk}-\bar{Y}_{\cdots}=(Y_{ijk}-\bar{Y}_{ij\cdot})+(\bar{Y}_{ij\cdot}-\bar{Y}_{\cdots})=(Y_{ijk}-\bar{Y}_{ij\cdot})+(\bar{Y}_{i\cdot\cdot}-\bar{Y}_{\cdots})+(\bar{Y}_{\cdot j\cdot}-\bar{Y}_{\cdots})+(\bar{Y}_{ij\cdot}-\bar{Y}_{i\cdot\cdot}-\bar{Y}_{\cdot j\cdot}-\bar{Y}_{\cdots})$. SST = SSE + SSA + SSB + SSAB

		B			
		$j=1$	$j=2$	$j=3$	Row means
A	$i=1$	$\mu_{11} = \mu+\alpha_1+\beta_1+(\alpha\beta)_{11}$	$\mu_{12} = \mu+\alpha_1+\beta_2+(\alpha\beta)_{12}$	$\mu_{13} = \mu+\alpha_1+\beta_3+(\alpha\beta)_{13}$	$\mu_{1\cdot} = \mu+\alpha_1$
	$i=2$	$\mu_{21} = \mu+\alpha_2+\beta_1+(\alpha\beta)_{21}$	$\mu_{22} = \mu+\alpha_2+\beta_2+(\alpha\beta)_{22}$	$\mu_{23} = \mu+\alpha_2+\beta_3+(\alpha\beta)_{23}$	$\mu_{2\cdot} = \mu+\alpha_2$
	Col means	$\mu_{\cdot 1} = \mu+\beta_1$	$\mu_{\cdot 2} = \mu+\beta_2$	$\mu_{\cdot 3} = \mu+\beta_3$	μ

Interaction effects:

$(\alpha\beta)_{11} = \mu_{11}-(\mu_{1\cdot}+\mu_{\cdot 1}-\mu)$	$(\alpha\beta)_{12} = \mu_{12}-(\mu_{1\cdot}+\mu_{\cdot 2}-\mu)$	$(\alpha\beta)_{13} = \mu_{13}-(\mu_{1\cdot}+\mu_{\cdot 3}-\mu)$
$(\alpha\beta)_{21} = \mu_{21}-(\mu_{2\cdot}+\mu_{\cdot 1}-\mu)$	$(\alpha\beta)_{22} = \mu_{22}-(\mu_{2\cdot}+\mu_{\cdot 2}-\mu)$	$(\alpha\beta)_{23} = \mu_{23}-(\mu_{2\cdot}+\mu_{\cdot 3}-\mu)$

Source of variation	df	Sum of squares	Mean square	F
Factor A	$a-1$	$SSA = \sum_i\sum_j\sum_k(\bar{Y}_{i\cdot\cdot}-\bar{Y}_{\cdot\cdot\cdot})^2$	$MSA = \frac{SSA}{a-1}$	$F_A = \frac{MSA}{MSE}$
Factor B	$b-1$	$SSB = \sum_i\sum_j\sum_k(\bar{Y}_{\cdot j\cdot}-\bar{Y}_{\cdot\cdot\cdot})^2$	$MSB = \frac{SSB}{b-1}$	$F_B = \frac{MSB}{MSE}$
Interaction AB	$(a-1)(b-1)$	$SSAB = \sum_i\sum_j\sum_k(\bar{Y}_{ij\cdot}-\bar{Y}_{i\cdot\cdot}-\bar{Y}_{\cdot j\cdot}+\bar{Y}_{\cdot\cdot\cdot})^2$	$MSAB = \frac{SSAB}{(a-1)(b-1)}$	$F_{AB} = \frac{MSAB}{MSE}$
Error	$N-ab$	$SSE = \sum_i\sum_j\sum_k(Y_{ijk}-\bar{Y}_{ij\cdot})^2$	$MSE = \frac{SSE}{N-ab}$	
Total	$N-1$	$SSTOT = \sum_i\sum_j\sum_k(Y_{ijk}-\bar{Y}_{\cdot\cdot\cdot})^2$		

  - Inference. Test for main effects same as above. Test for interactions: $H_0^{AB}: (\alpha\beta)_{ij} = 0\ \forall i,j$. F test: $F = \frac{SSAB/[(a-1)(b-1)]}{SSE/(N-ab)} = \frac{MSAB}{MSE}$
  - Simple main effects: effects of a factor at a specific level of the other factor

- Means plots. X = one factor, Y = outcome, split by other factor. Parallel lines indicate additive effects, nonparallel indicate interactions.
- Unequal cell sizes. Factors are now correlated, sums of squares are not orthogonal, tests of effects are not independent, and SST $\neq$ SSA + SSB + SSAB. Use adjusted means: the marginal mean as if the design were balanced.

2.3.3 ANCOVA: Analysis of covariance

Add continuous concomitant variables to increase statistical power (by reducing error variance) and adjust for group differences.

- Model (single factor): $y_{ij} = \mu+\tau_i+\gamma X_{ij}+\varepsilon_{ij},\ \varepsilon_{ij} \sim N(0,\sigma^2)$, with $\sum_i \tau_i = 0$. Group means are interpreted when $X_{ij} = 0$. Can also use mean centering: $y_{ij} = \mu+\tau_i+\gamma(X_{ij}-\bar{X}_{\cdot\cdot})+\varepsilon_{ij}$. In general can have additional factors or multiple concomitant variables.
- Reduces error variance by modeling additional variation in the outcome

- In pre-post designs, can use baseline score as a covariate.
- Additional model assumption: equality of slopes for different treatments = parallel regression functions. Can check by testing interaction terms.
- Possible ANCOVA problems: little gain in precision due to correlation between concomitant and factor variables, or hidden extrapolation because of nonoverlapping ranges of the concomitant variable among groups.

3 Biostat 200C

3.1 Generalized linear models

- Modeling non-normal responses as a function of covariates:

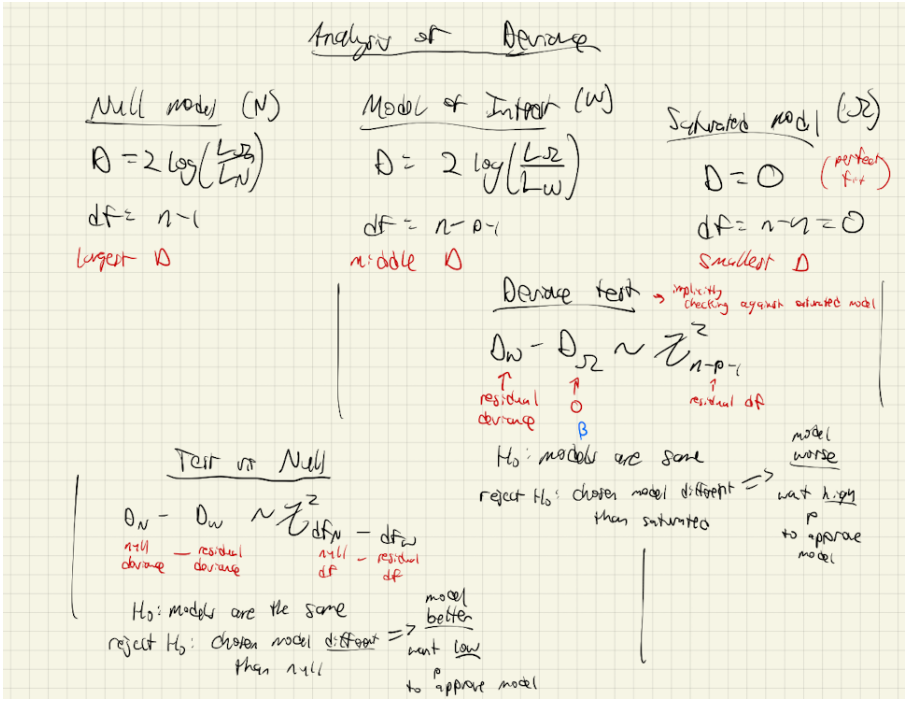
$$\begin{aligned} & E(Y) = \mu = g^{-1}(\eta) = g^{-1}(x^T\beta) \\ \iff \quad & g(\mu) = \eta \qquad\quad = x^T\beta \end{aligned}$$

4 where $g(.)$ is a link function relating the (conditional) mean to the linear predictor, and $\beta \in \mathbb{R}^{q+1}$. Note a link function is a function of the conditional mean, and the *inverse* link function is a function of the linear predictor. Note GLMs have no additive error term.

- Deviance. Check goodness of model fit. For a model with p parameters, deviance is the LRT statistic versus the saturated model with $m \leq n$ parameters having MLE b_{max}:

$$D = 2\log\frac{L(b_{max})}{L(b)} = 2[\ell(b_{max}) - \ell(b)] \sim \chi^2_{m-p}$$

under H_0. (Note that this formula includes the negative sign in the log and puts b_{max} in the numerator, equivalent to 202b.)



  - Goodness of fit test is thought of as current model vs. the saturated model. A high p-value, failing to reject H_0, indicates no evidence the models are different, indicating good model fit. We WANT our model to fit as well as the saturated model.
  - Null test. Instead of comparing current vs. saturated, compare current vs. null (intercept) model. A low p-value, rejecting H_0, indicates the current model is different from the null, indicating improvement over the null model. We WANT to be different from the null model.

  - Test versus another candidate model.

$$D_{small} - D_{large} \sim \chi^2_{df_S - df_L}$$

    Rejecting H_0 indicates including the additional parameters is important,
  - Residual deviance: deviance of current model
  - Null deviance: deviance of null model with just an intercept
  - Assuming the null model, $D_S - D_L \sim \chi^2_{l-s}$
  - If a dispersion parameter is estimated (by $\hat{\phi} = \chi^2_{pearson}/(n-p)$), then we use an F test:

$$\frac{(D_S - D_L)/(df_S - df_L)}{\hat{\phi}} \sim F_{df_S - df_L, n-p}$$

- Residuals + diagnostics
  - Raw residuals: $y - \hat{p}$
  - Deviance residuals: $r_i^D = sgn(y_i - \hat{p_i})\sqrt{d_i}$ s.t. $\sum_i d_i^2 = D$. E.g. for logistic regression $d_i = sgn(y_i - \hat{p_i})\sqrt{-2[y_i \log \hat{p}_i + (1-y_i)\log(1-\hat{p}_i)]}$.
  - Pearson residuals: $r_i^P = \frac{y_i - n_i \hat{p}_i}{\sqrt{Var(\hat{y}_i)}}$ s.t. $\sum_i (r_i^P)^2 = \chi^2_{Pearson}$ for binomial GLM.
  - Studentized residuals: $r_i^{SD} = \frac{r_i^D}{\sqrt{\hat{\phi}(1-h_i)}}$ using the hat values
  - Half-normal plot: plotting hat values vs. quantiles of half-normal distribution. Identify high leverage cases/outliers in predictor space.
  - Cook's distance vs. half-normal plot. Identify influential cases.
- Model selection
  - AIC = deviance + 2q
  - Lasso: $\arg\min_\beta \frac{1}{n}\ell(\beta) + \lambda \sum_j |\beta_j|$
- Overdispersion. Present when variance of data is greater than that indicated by distribution of model. Can relax distribution assumptions by adding an overdispersion parameter to variance function. Can also use sandwich estimation for the estimator of $Var\hat{\beta}$, or robust estimation.
- Exponential family distribution. θ = canonical parameter, ϕ = dispersion parameter:

$$f(y|\theta, \phi) = \exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right]$$

  - Moments:

$$EY = \mu = b'(\theta)$$
$$VarY = \sigma^2 = b''(\theta)a(\phi)$$

  - Canonical link is g such that $\eta = g(\mu) = \theta$
- Fisher scoring algorithm. GLMS are fit using MLE, with log-likelihood maximized using Newton-Raphson (aka IRWLS):

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} + s[-\nabla^2 \ell(\boldsymbol{\beta}^{(t)})]^{-1} \nabla \ell(\boldsymbol{\beta}^{(t)})$$

3.1.1 GLM Examples

n observations, q predictors plus 1 intercept

- Binary response:
  - $Y_i = \begin{cases} 1 & \text{with probability } p_i \\ 0 & \text{with probability } 1 - p_i \end{cases}$
  - Link function: $\eta = g(p) = \log(\frac{p}{1-p})$ (logit). Inverse link function (logistic):
$$p = g^{-1}(\eta) = \frac{e^\eta}{1 + e^\eta}$$
  - Log-likelihood:
$$\begin{aligned} \ell(\beta) &= \sum_i \log[p_i^{y_i}(1 - p_i)^{1-y_i}] \\ &= \sum_i [y_i \cdot x_i^T \beta - \log(1 + e^{x_i^T \beta})] \end{aligned}$$
  - Interpretation: additive change in log odds, or multiplicative change in odds of outcome
  - Deviance:
$$D = -2 \sum_i [y_i \log \hat{p}_i + (1 - y_i) \log(1 - \hat{p}_i)]$$
    Note can't use standard goodness of fit test vs. χ^2_{n-q-1} because one observation per covariate combination "bin", but can conduct test vs. null.
  - Goodness of fit using Hosmer-Lemeshow. Make J bins of linear predictors, then observed bin proportion should be close to average predicted probability: $\chi^2_{HL} = \sum_j \frac{(y_i - m_j \hat{p}_j)^2}{m_j \hat{p}_j (1 - \hat{p}_j)} \sim \chi^2_{J-1}$
  - ROC curve: plot sensitivty vs. 1-specificity = true positive rate vs. false positive rate
  - Choice of link function corresponds to different latent variable T with $Y = 1$ if $T \leq t$, 0 otherwise
    - Logit link: $\eta = \log \frac{p}{1-p} \iff$ Logistic latent variable
    - Probit link: $\eta = \Phi^{-1}(p) \iff$ Normal latent variable
    - Complmentary log-log: $\eta = \log(-\log(1 - p)) \iff$ Gumbel latent variable
    - Cauchit link: $\eta = \tan((p - 1/2)\pi) \iff$ Cauchy latent variable
  - Prospective sampling: predictors are fixed and outcome is observed, possible to estimate intercept $\implies$ possible to estimate risk. Retrospective sampling: outcome fixed and predictors are observed, can't estimate intercept $\implies$ can't estimate risk, only relative effect (odds ratios)
  - Effective dose 50 (ED50): dose corresponding to a 50% chance of success. Solution to $p = g^{-1}(x^T \beta) = 0.5$
  - Matched case-control study (1:M design). Fit conditional logistic regression, with form of conditional likelihood function identical to that for Cox proportional hazards:
$$\begin{aligned} L(\beta) &= \prod_{j=1}^n P\left(Y_{0j} = 1, Y_{1j} = \cdots = Y_{mj} = 0 | \sum_{i=0}^M Y_{ij} = 1\right) \\ &= \prod_{j=1}^n \frac{1}{1 + \sum_{i=1}^M \exp(x_{ij} - x_{0j})^T \beta} \end{aligned}$$

  - Perfectly separable data: when classes are able to be perfectly split by a hyperplane. Causes estimation issues.

- Binomial response. Binary is a special case with $m_i = 1 \; \forall i$
  - $P(Y_i = y_i) = \binom{m_i}{y_i} p_i^{y_i}(1-p_i)^{m_i - y_i}$ for batch sizes m_i and success probability p_i
  - Logistic inverse link: $p_i = \frac{e^{\eta_i}}{1+e^{\eta_i}}$
  - Log-likelihood:

$$\begin{aligned}\ell(\beta) &= \sum_i [y_i \log p_i + (m_i - y_i)\log(1-p_i) + \log \binom{m_i}{y_i}] \\ &= \sum_i [y_i \cdot x_i^T \beta - m_i \log(1 + e^{x_i^T \beta}) + \log \binom{m_i}{y_i}]\end{aligned}$$

  - Deviance:

$$D = 2\sum_i y_i \log(\frac{y_i}{\hat{y}_i}) + (m_i - y_i)\log\frac{m_i - y_i}{m_i - \hat{y}_i} \sim \chi^2_{n-q-1}$$

    Null approximation holds if $m_i \geq 5$.
  - Goodness of fit using Pearson χ^2. Alternative to deviance test: $\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i} = \sum_i \frac{(y_i - n_j\hat{p}_j)^2}{n_j\hat{p}_j(1-\hat{p}_j)} \sim \chi^2_{n-q-1}$
  - Overdispersion. Add overdispersion parameter s.t. $Var(Y) = \sigma^2 np(1-p)$. Estimate of the parameter $\hat{\sigma}^2 = \frac{\chi^2_{Pearson}}{n-q-1}$. Parameter estimates are unchanged, but $Var(\hat{\beta}) = \hat{\sigma}^2(X^T\hat{W}X)^{-1}$. Perform analysis of deviance using F test:

$$F = \frac{(D_{small} - D_{large})/(df_{small} - dflarge)}{\hat{\sigma}^2} \sim F_{df_{small} - df_{large}, n-q-1}$$

  - Quasi-binomial model. Also deals with over/under-dispersion. Assumes $E(Y_i) = \mu_i, Var(Y_i) = \phi V(\mu_i)$. Maximizing a log quasi-likelihood.
  - Beta regression. Models proportions directly using beta distribution.

- Poisson regression
  - $P(Y_i = y_i) = e^{-\mu_i}\frac{\mu_i^{y_i}}{y_i!}$
  - Exponential inverse link (log link): $\mu_i = e^{\eta_i} \iff \eta_i = \log(\mu_i)$
  - Log-likelihood:

$$\begin{aligned}\ell(\beta) &= \sum_i y_i \log\mu_i - \mu_i - \log y_i! \\ &= \sum_i y_i \cdot x_i^T\beta - e^{x_i^T\beta} - \log y_i!\end{aligned}$$

  - Interpretation: additive change in log poisson mean, or multiplicative change in poisson mean, or (additive/multiplicative) change in poisson rate
  - Deviance (G-statistic):

$$2\sum_i y_i \log(\frac{y_i}{\hat{\mu}_i}) \sim \chi^2_{n-p}$$

    Can also use Pearson Chi-square: $\chi^2_{Pearson} = \sum_i \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i} \sim \chi^2_{n-p}$

  - Overdispersion. Variance function now $Var(Y_i) = \phi\mu_i$, estimate dispersion parameter by again Pearson chi-square divided by degrees of freedom.
  - Quasi-Poisson model. Assumes $E(Y_i) = \mu_i, Var(Y_i) = \phi\mu_i$.
  - Negative binomial regression. Assumes Y follows negative binomial distribution, also uses log link.
  - Zero-inflated count models. Separately models probabilities associated with zeros using a Bernoulli outcome, but does it in reverse ways.
    * Hurdle model: f_1 models probability of observing nonzero (Bernoulli), f_2 is a zero-truncated Poisson.

$$P(Y = 0) = f_1(0)$$
$$P(Y = j) = \frac{1 - f_1(0)}{1 - f_2(0)} f_2(j), \quad j > 0$$

    * Zero-inflated poisson: ϕ models models probability of always observing zero (Bernoulli), f is Poisson.

$$P(Y = 0) = \phi + (1 - \phi)f(0)$$
$$P(Y = j) = (1 - \phi)f(j), \quad j > 0$$

- Multinomial data. Have cases for nominal and ordinal data with J levels.
  - Multinomal logit. Nominal data. Considering log odds of belonging to group j versus reference group 0. Each model has own intercept and slope terms, so (q+1)(J-1) total parameters

$$\log\left(\frac{p_j}{p_0}\right) = \beta_{0j} + \beta_{1j}X_1 + \cdots + \beta_{qj}X_q$$

  - Proportional odds model. Ordinal data. Assumes the change in odds of moving from one category to next level up is same regardless of current category. The intercept is the probability of moving from one specific level to another when all predictors are equal to 0. So slopes are the same, but each level gets own intercept (minus reference). So q + J - 1 parameters.

$$\log\left(\frac{P(Y_i \leq j)}{P(Y_i > j)}\right) = \log\left(\frac{P(Y_i \leq j)}{1 - P(Y_i \leq j)}\right) = \theta_{0,j} - \beta_1 X_1 - \cdots - \beta_q X_q$$

    * Using probit link gives ordered probit model
    * Using complementary log-log link gives proportional hazards model

- 2x2 Contingency tables. Different sampling schemes give rise to different models. Example: N items with variables A (levels 1-2), B (levels 1-2)
  - Poisson: observe N items over time, then cross-classify by variables A and B
  - Multinomial: sample N items, then cross-classify by variables A and B
  - Binomial: sample K items with feature A1, N-K items with feature A2, then cross-classify by variable B
  - Hypergeometric: sample K items with feature A1, N-K itms with feature A1, with the number of items from B1 and B2 fixed (fixed marginal totals).

- 3-way Contingency tables. 4 different association structures, each give rise to different models

- Mutual independence model $\implies$ Poisson model, main effects only

$$p_{ijk} = p_i p_j p_k \implies EY_{ijk} = np_{ijk}$$
$$\log EY_{ijk} = \log n + \log p_i + \log p_j + \log p_k$$

- Joint independence model $\implies$ Poisson model, one interaction term

$$p_{ijk} = p_{ij} p_k$$
$$\log EY_{ijk} = \log n + \log p_{ij} + \log p_k$$

- Conditional independence model $\implies$ Poisson model, two interaction terms

$$p_{ijk} = p_{ij|k} p_k = p_{i|k} p_{j|k} p_k = \frac{p_{ik} p_{jk}}{p_k}$$
$$\log EY_{ijk} = \log n + \log p_{ik} + \log p_{jk} - \log p_k$$

- Uniform association model $\implies$ Poisson model, all 2-way interactions

$$\log EY_{ijk} = \log n + \sum_{x=i,j,k} \log p_x + \sum_{y=ij,jk,ik} \log p_y$$

3.2 Mixed models

- Mixed effects model:

$$Y = X\beta + Z\gamma + \varepsilon$$

where $X \in \mathbb{R}^{n \times p}$ the design matrix for the fixed effects $\beta \in \mathbb{R}^p$, $Z \in \mathbb{R}^{n \times q}$ the design matrix for the fixed effects $\gamma \in \mathbb{R}^q$, and $\varepsilon \sim N(0_n, \sigma^2 I), \gamma \sim N(0_q, \Sigma)$, with $\varepsilon,\ \gamma$ independent. So:

$$Y \sim N(X\beta, Z\Sigma Z^T + \sigma^2 I)$$

- Fit using maximum likelihood. Likelihood is derived from multivariate normal. Results in biased variance component parameter estimates $\hat{\sigma}^2, \hat{\Sigma}$
- REML. Reduces the above bias from MLE. Work with a transformation of Y, $K^T Y$ using a basis K of the left null space of the fixed effect design matrix, $\mathcal{N}(X^T) = \mathcal{C}(X)^\perp$. First estimate variance component parameters using MLE and transformed data. Then use general least squares to estimate fixed effects.
- Inference
  - Fixed effects. For nested models with the same random effects, use LRT (fitted using MLE). Prefer to use Kenward-Roger adjusted F test (fitted using REML) using adjusted df.
  - Variance component parameters.
    - Parametric bootstrap. Since conventional null distribution can be wrong (boundary condition, nonnegative variance). Generate new Y from fitted null model, calculate LRT vs. candidate model with random effects. Proportion of replicates with LRT stats larger than observed LRT is the p-value for the test. Note this also works for fixed effects
    - Also can use exact LRT, RLRT

- Random effect estimation. From a Bayesian POV, we have likelihood $y|\gamma \sim N(X\beta + Z\gamma, \sigma^2 I)$ and prior $\gamma \sim N(0_q, \Sigma)$. Can get the posterior, which is multivariate normal. Can use the posterior mean $E(\gamma|y)$ to estimate random effects.
- Prediction. Obtain best linear unbiased predictors.
- Diagnostics. Use residuals calculated using predicted random effects, serve as estimates of ε. Standard plots: QQ, residual vs. fitted.
- Randomized block designs. Treat blocking variable as a random effect.
- Longitudinal models. Can have random intercept providing each subject an intercept at baseline, and random slope providing each subject a different effect of time on the outcome.
  - Can also have nesting of repeated measures within subjects.
- Generalized Linear Mixed Models
  - Have $f(y_i|\theta_i, \phi)$ from the exponential family, with canonical link such that $\theta_i = g(\mu_i) = \eta_i$, with
    $$\eta_i = x_i^T \beta + z_i^T \gamma$$
    where β are fixed effects, and γ are random effects.
  - Estimation and inference.
    - Maximum likelihood estimation. Needs numerical integration, for example Gauss-Hermite quadrature or Laplace approximation.
    - Bayesian methods
    - Penalized quasi-likelihood. Adapt working response in IRWLS to mixed effects model.
    - Generalized estimation equations (GEE). Below.
  - Generalized estimation equations (GEE). Generalizes quasi-likelihood approach. Model:
    $$\begin{aligned} \mathbb{E}\mathbf{Y}_i &= \boldsymbol{\mu}_i \\ g(\boldsymbol{\mu}_i) &= \mathbf{X}_i\boldsymbol{\beta} \\ \mathbf{V}_i &= \mathrm{Var}(\mathbf{Y}_i) = \phi \mathbf{A}_i^{1/2} \mathbf{R}_i(\boldsymbol{\alpha}) \mathbf{A}_i^{1/2}, \end{aligned}$$
    where $A_i = diag(a(\mu))$ captures individual variances, $R_i(\alpha)$ is a working correlation matrix. Options for the working correlation matrix include compound symmetry/exchangeable correlation, autoregression, or unstructured.
    Estimation equation:
    $$\sum_i [D_{\boldsymbol{\beta}} \boldsymbol{\mu}_i(\boldsymbol{\beta})]^T \mathbf{V}_i^{-1} (\mathbf{Y}_i - \boldsymbol{\mu}_i) = \mathbf{0}.$$
    GEE estimates are half the size of those from GLMM, since GEE is marginalized over all individuals with sets of covariates, whereas GLMM is at the individual level.

3.3 Other topics

- Survival analysis
  - Characteristics: survival times are non-negative and usually right skewed, and subjects may be censored.
  - Important functions

* Survivor function: probability of survival beyond time y

$$S(y) = P(Y > y) = 1 - F(y)$$

where F is the cdf of the failure time distribution.

* Hazard function: probability of event in an infinitesimal interval, given survival to time y

$$\begin{aligned} h(y) &= \lim_{\Delta y \to 0} \frac{\mathbb{P}(y < Y \leq y + \Delta y \mid Y > y)}{\Delta y} \\ &= \frac{f(y)}{S(y)} \\ &= -\frac{d}{dy} \log S(y) \end{aligned}$$

* Have relationship between survival function and hazard function:

$$\begin{aligned} S(y) &= e^{-H(y)} = e^{-\int_0^y h(t)dt} \\ H(y) &= -\log S(y) \end{aligned}$$

where $H(y)$ is the cumulative/integrated hazard function.

– Exponential model. Specify $f(y;\theta) = \theta e^{-\theta y}$. Gives memoryless property: hazard function is $h(y;\theta) = \theta$, doesn't depend on y.

– Proportional hazards model. Using log link: $\theta = e^{x^T\beta} \implies h(y;\beta) = e^{x^T\beta}$, with a constant hazard ratio for a unit change in x_k.

Proportional hazards models have the form

$$h_1(y) = h_0(y)e^{x^T\beta}$$

with

$$\log H_1(y) = \log H_0(y) + x^T\beta$$

– Weibull distribution. Use survival time distribution $f(y;\lambda,\theta) = \frac{\lambda y^{y-1}}{\theta^\lambda} e^{(y/\theta)^\lambda}$, $y \geq 0, \lambda > 0, \theta > 0$. Let $\phi = \theta^{-\lambda}$. The exponential model is a special case with $\lambda = 1$. Have hazard function:

$$h(y;\lambda,\phi) = \lambda\phi y^{\lambda-1}$$

If $\lambda > 1$, have increasing Weibull/accelerated failure time model, with increasing probability of death over time. If $\lambda < 1$, have decreasing Weibull model, and probability of death decreases.

– Kaplan-Meier estimate. Nonparametric estimate of survival function $\hat{S}(y) = \hat{P}(Y > y)$:

$$\begin{aligned} \hat{S}(y_k) &= \hat{P}(Y > y_{(k)}) \\ &= \prod_{j=1}^{k} \hat{P}(Y > y_{(j)} | Y > y_{(j-1)}) \\ &= \prod_{j=1}^{k} \left(\frac{n_j - d_j}{n_j} \right) \end{aligned}$$

where n_j is the number alive just before time $y_{(j)}$, and d_j is the number of events at time $y_{(j)}$

  - Estimation. Using likelihood $L = \prod_{j=1}^{n} f(y_j)^{\delta_j} S(y_j)^{1-\delta_j}$, use MLE. Note for the exponential model, the log-likelihood resembles the log-likelihood for a Poisson model for δ_j with $\mu_j = \theta_j y_j$, and can estimate this using Poisson regression with an offset.
  - Diagnostics.
    * Plot (for exponential model) $-\log(\hat{S}(y))$ against y, should be linear. Analogous for Weibull.
    * Check proportional hazards assumption. Stratifying above plot by covariates should result in parallel lines.
    * Residuals. Have Cox-Snell residuals, Martingale residuals, and Deviance residuals.

- Nonparametric regression. In general, for the model $y_i = f(x_i) + \varepsilon$, we can just assume f is from some smooth family of functions, instead of belonging to a parametric family $f(x|\beta)$
  - Kernel estimators. Weight each observation using a kernel function.
  - Smoothing splines
  - Regression splines
  - Local polynomials
  - Wavelets

- Generalized Additive Models (GAM). Nonparametric regression becomes impractical for many predictors.
  - Additive models. For p continuous predictors x_j, and a vector of categorical predictors z, have the model:
    $$y = \beta_0 + \sum_{j=1}^{p} f_j(x_j) + z^T\gamma + \varepsilon$$
    where f_j are smooth functions for each predictor separately. Fit using a backfitting algorithm.
  - Generalized additive models. Now using a link function with the systematic component $\eta = \beta_0 + \sum_{j=1}^{p} f_j(x_j)$
  - Generalized additive mixed models.

- Trees
  - Divide predictor space into non-overlapping boxes whcih minimze RSS, where the prediction in each box $\hat{y}_{R_j}$ is the mean of the response for all training observations in R_j. Split using a top-down, greedy approach.
  - Grow a large tree, then use cost-complexity pruning to avoid overfitting. Use cross-validation to choose pruning tuning parameter.
  - Also have classification trees.
  - Bagging/bootstrap aggregation. Reduces the variance of the estimate by creating bootstrap samples from the training set, averaging predictors for final estimate. Estimate test error using out-of-bag (OOB) error estimation.
  - Random forests. Decorrelate trees by using bootstrapped samples and random subsets of the predictors, usually $m = \sqrt{p}$ many.
  - Boosting. Grow trees sequentially, with each subsequent tree using information from previous trees and trying to fit to residuals. Have tuning parameters number of trees, shrinkage parameter/learning rate, number of tree splits.

- Variable importance. Average decrease in RSS due to splits over a given predictor across all bagged trees. For classification, add total decrease in Gini index over a predictor across all bagged trees.

4 Biostat 202A

- Probability
  - Experiment $\implies$ Sample space $\mathcal{S} \implies$ Events $A \subset \mathcal{S}$.
  - Set operations on events: $\subset, =, \cup, \cap, A^C$
  - σ-field/algebra $\mathcal{F}$ satisfying:
    1. $\varnothing \in \mathcal{F}$
    2. Closed under countable unions: $\cup_{i=1}^{k} A_i \in \mathcal{F}$ if $A_i \in \mathcal{F}\ \forall i$
    3. Closed under complementation: $A \in \mathcal{F} \implies A^C \in \mathcal{F}$

    Useful example: $\mathcal{B} = \{$all possible subsets of S, including $\varnothing, S\}$ is a σ-algebra, and if $\mathcal{I} =$ collection of all open intervals (a, b) $s.t.$ $a < b$, then $\mathcal{B}_0 = \cap\{\mathcal{B} | \mathcal{I} \subset \mathcal{B}\}$ the Borel algebra, is the smallest sigma-algebra that contains the open sets.
  - Probability measure $P : \mathcal{F} \to \mathbb{R}$ satisyfing:
    1. $P(A) \geq 0$
    2. Countable additivity: $P(\cup_i A_i) = \sum_i P(A_i)$ for disjoint $A_i s$
    3. P(S) = 1
  - Probability space $(\mathcal{S}, \mathcal{F}, P)$
- Discrete probability. The sample space is countable: $\mathcal{S} = \{s_1, s_2, \dots\}$. If we assign probabilities to each event $p_1, p_2, \dots$ $s.t. \sum p_i = 1$, then

$$P(A) = \sum_{s_i \in A} p_i$$

  is a probability function.
- Continuous probability. For a nonnegative function $f \geq 0$ with $\int_{-\infty}^{\infty} f(x)dx = 1$, then

$$P(A) = \int_A f(x)dx$$

  is a probability function.
- Counting
  - In the discrete case, for a discrete sample space with $p_i = 1/n\ \forall i$, then $P(A) = \frac{\#\text{ outcomes in A}}{\#\text{ outcomes in } \mathcal{S}}$
  - To arrange r subjects from n items:

	Without replacement	With replacement
Ordered (distinguishable)	$\frac{n!}{(n-r)!}$	n^r
Unordered (indistinguishable)	$\frac{n!}{r!(n-r)!} = \binom{n}{r}$	$\binom{n+r-1}{r}$

- Conditional probability
  - Definition:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \quad P(B) \geq 0$$

  - Independence:

$$\begin{aligned} P(A \cap B) &= P(A)P(B) \iff A \perp B \\ \iff P(A|B) &= P(A) \end{aligned}$$

  - Bayes theorem:

$$P(A_j|B) = \frac{P(B|A_j)P(A_j)}{P(B)} = \frac{P(B|A_j)P(A_j)}{\sum_{i=1}^{k} P(B|A_i)P(A_i)}$$

    where $\{A_1, \dots A_k\}$ partition $\mathcal{S}$
  - Note conditional probability is its own probability function and induces its own probability space: $P(\cdot|B) \implies (B, B \cap \mathcal{F}, P_B(\cdot))$, where $P_B(\cdot) = \frac{P(A \cap B)}{P(B)}$

- Random variables. Functions $X : \mathcal{S} \to \mathbb{R}$ satisfying $X^{-1}(B) \in \mathcal{F}$ for $B \in \mathcal{B}_0$
  - Random vectors $X : \mathcal{S} \to \mathbb{R}^n$
  - X induces a new probability space: $(\mathcal{S}, \mathcal{F}, P) \to (\mathbb{R}, \mathcal{B}_0, P_X)$, where $P_X(A) = P(X^{-1}(A))$ for $A \in \mathcal{B}_0$
  - Probability function P_X has cumulative distribution function (cdf) $F_X(x) = P(X \leq x)$, with probability mass function (pmf) if the range of X is discrete, and probability density function (pdf) if the range of X is an interval.

    Cdfs satisfy:
    1. $\lim_{X \to -\infty} F(x) = 0, \quad \lim_{X \to \infty} F(x) = 1$
    2. $F(x)$ nondecreasing
    3. $F(x)$ right-continuous at any x_0: $\lim_{x \downarrow x_0} F(x) = F(x_0)$

- Transformations. $Y = g(X)$. Note whether g is one-to-one or not.

$$\begin{aligned}
F_Y(y) &= P(Y \leq y) \\
&= P(g(X) \leq y) \\
&= \begin{cases} P(X \leq g^{-1}(y)) = F_X(g^{-1}(y)) & \text{if g increasing} \\ P(X \geq g^{-1}(y)) = 1 - F_X(g^{-1}(y)) & \text{if g decreasing} \end{cases} \\
f_Y(y) &= \frac{\partial}{\partial y} \int_{-\infty}^{g^{-1}(y)} f_X(x)dx \\
&= f_X(g^{-1}(y)) \left| \frac{dx}{dy} \right| \\
&= \sum_{i=1}^{k} f_X(g_i^{-1}(y)) \left| \frac{d}{dy}\{g_i^{-1}(y)\} \right| \quad \text{if g is not monotone over all of } \mathcal{X}
\end{aligned}$$

  where $\frac{dx}{dy} = \frac{d}{dy}\{g^{-1}(y)\}$.
  - If $g(.)$ is non-monotonic, partition $\mathcal{X}$. If the range $\mathcal{Y}$ of g on each partition of $\mathcal{X}$ is not the same, partition $\mathcal{Y}$ as well.
  - If multivariate: $g = (g_1, \dots, g_n)$, and $f_Y(y_1, \dots, y_n) = f_X(w_1(y_1, \dots, y_n), \dots, w_n(y_1, \dots, y_n))|J|$, where $w_i = g_i^{-1}(y)$
  - How to find the bounds of Y, defining $\mathcal{Y}$? Do this *first*.
    * The boundaries of $\mathcal{X}$ define the boundaries of $\mathcal{Y} \implies$ plug in the bounds to find expressions for the boundaries of $\mathcal{Y}$, if for example we want to integrate over them.
  - Note that if $X \sim Unif(0, 1)$, and F is a cdf, then $Y = F^{-1}(x) \sim F$.

- Expectation, Variance

- Moments: $E(X^m)$, moment generating function (mgf): $E(e^{tz})$. For a transformed variable, $M_{aX+b}(t) = e^{bt}M_X(at)$
- Iterated expectations and conditional expectation/variance:

$$\begin{aligned} E(E(Y|X)) &= E(Y) \\ Var(Y) &= Var(E(Y|X)) + E(Var(Y|X)) \end{aligned}$$

  Noe that conditional expectation and conditional variance $Y|X$ are functions of X.
- Covariance, correlation. How to prove $|p| \leq 1$ using Hölder's inequality.
- Inequalities
  - Chebyshev:

$$\begin{aligned} P(g(x) \geq r) &\leq \frac{Eg(x)}{r} \\ P(|X - \mu| \geq k\sigma) &\leq \frac{1}{k^2} \end{aligned}$$

    Using Chebyshev to prove the WLLN
  - Lemma: If $\frac{1}{p} + \frac{1}{q} = 1$, then

$$\frac{1}{p}a^p + \frac{1}{q}b^q \geq ab$$

  - Hölder:

$$\begin{aligned} |E(XY)| &\leq E|XY| \\ &\leq (E|X|^p)^{1/p}(E|Y|^q)^{1/q} \end{aligned}$$

  - Liapunov: For $1 < r < s < \infty$,

$$(E|X|^r)^{1/r} \leq (E|X|^s)^{1/s} \quad \text{(lower moments bounded by higher moments)} \tag{1}$$

$$E|X| \leq (E|X|^p)^{1/p}, \quad 1 < p < \infty \tag{2}$$

  - Minkowski: For $1 < p < \infty$,

$$(E|X + Y|^p)^{1/p} \leq (E|X|^p)^{1/p} + (E|Y|^p)^{1/p}$$

  - Jensen: For g a convex function,

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

    Using this to prove that harmonic mean $\leq$ geometric mean $\leq$ arithmetic mean
- Multivariate normal distribution
  - If $\mathbf{X} = \begin{bmatrix} \mathbf{X_1} \\ \mathbf{X_2} \end{bmatrix} \sim N_n(\mu, \mathbf{\Sigma})$, where $\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$, $\mathbf{\Sigma} = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$, and $dim(X_1) = m < n$ then:
    - If off-diagonal entries in covariance matrix are 0, then each univariate distribution is independent: $X_1 \perp X_2 \iff \Sigma_{12} = 0$
    - $\mathbf{X_1} \sim N_m(\mu_\mathbf{1}, \mathbf{\Sigma_{11}})$. Same for $\mathbf{X_2}$.
    - For $A \in \mathbb{R}^{m \times n}, m < n$,

$$Y = AX + b \sim N(A\mu + b, A\Sigma A^T)$$

    - The conditional distribution:

$$X_1|X_2 \sim N_m(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})$$

4.1 Distributions to know

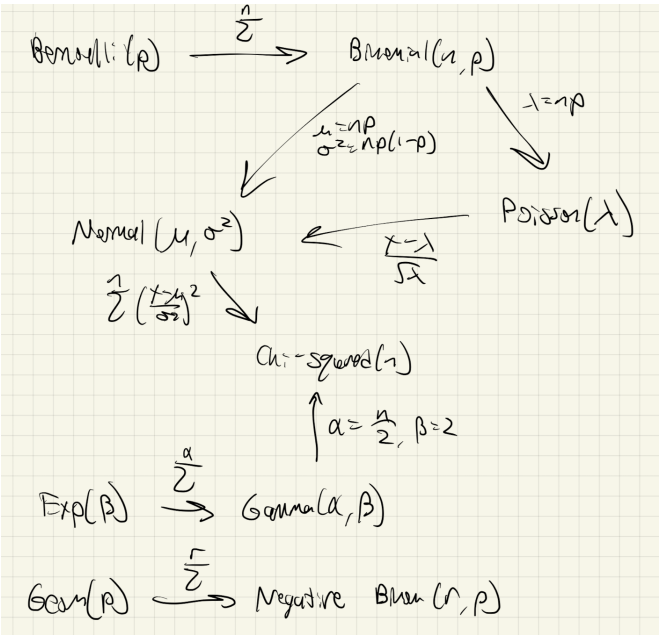
Know the pdf/pmf, mean, variance, mgf, domain. Know interpretations and relationships. For hard to remember pdfs, remember the kernel.

4.1.1 Continuous distributions

- Uniform (a, b)
- (Logistic (μ, β))
- Normal (μ, σ^2)
- Exponential (β)
- (Cauchy (θ, σ))
- Gamma (α, β)
- Chi-squared (n)
- Multivariate Normal (μ, Σ)
- T (p) (202B)
- F (p, q) (202B)
- Beta (α, β) (202B)

4.1.2 Discrete distributions

- Uniform $(1, N)$
- Bernoulli (p)
- Binomial (n, p)
- Geometric (p)
- Negative binomial (r, p)
- Multinomial $(n, p_1, \ldots, p_k)$. Including order statistics (202B).
- Hypergeometric (N, M, K)
- Poisson (λ), with Poisson postulates


Bernoulli(p)
$\sum$
Binomial(n, p)
$\lambda = np$
$\mu = np$
$\sigma^2 = np(1-p)$
Normal (μ, σ^2)
Poisson(λ)
$\frac{X-\lambda}{\sqrt{\lambda}}$
$\sum \left(\frac{X-\mu}{\sigma}\right)^2$
Chi-squared(n)
$\alpha = \frac{n}{2}, \beta = 2$
Exp(β)
$\sum^{\alpha}$
Gamma(α, β)
Geom(p)
$\sum^{r}$
Negative Binom(r, p)

5 Biostat 202B

- T, F, Beta distributions
  - $T = \frac{U}{\sqrt{V/p}} \sim t_p$, where $U \sim N(0,1),\ V \sim \chi^2_p,\ U \perp V$
  - $F = \frac{U/p}{V/q} \sim F_{p,q}$, where $U \sim \chi^2_p,\ V \sim \chi^2_q,\ U \perp V$
  - Beta(α, β), between 0 and 1
- Random samples
  - $X_1, \ldots, X_n$ is SRS from F if iid with cdf F
  - Statistics are functions of random samples: $T = T(x_1, \ldots, x_n)$, with distribution called a sampling distribution
  - $\bar{X}, S^2$ properties:
    1. $E(\bar{X}) = \mu = E(X_i)$
    2. $Var(\bar{X}) = \sigma^2/n$
    3. $E(S^2) = \sigma^2$
    4. $M_{\bar{X}}(t) = [M_X(\frac{t}{n})]^n$
- Samples from a Normal distribution: $X_i \sim N(\mu, \sigma^2)\ \forall i$
  - Statistic properties:
    1. $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$. Prove using mgf.
    2. $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$. Prove using sum of n-1 deviations.
    3. $\bar{X} \perp S^2$
  - Inference on μ.
    - Point estimate: $\bar{X}$
    - For a $1-\alpha$ CI: $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1) \implies$

$$P\left(-Z_{1-\frac{\alpha}{2}} \leq \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq Z_{1-\frac{\alpha}{2}}\right) = 1-\alpha$$
$$\implies \left[\bar{X} \pm Z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right]$$

    - If variance unknown:

$$\frac{\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1)S^2}{n\sigma^2}/(n-1)}} \overset{N(0,1)}{\sqrt{\chi^2_{n-1}}} = \frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}$$
$$\implies \left[\bar{X} \pm t_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}}\right]$$

    - For two samples, analogous: $\bar{X}_1 - \bar{X}_2 \sim N(\mu_1 - \mu_2, \sigma^2/n_1 + \sigma^2/n_2)$. If variance unknown, use pooled $S_p^2 \sim \chi^2_{n_1+n_2-2}$.
  - Inference on σ^2
    - Point estimate: $S^2 = \frac{1}{n-1}\sum(x_i - \bar{X})^2$
    - A $1-\alpha$ CI: $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1} \implies$

$$P\left(\chi^2_{\alpha/2} \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi^2_{1-\alpha/2}\right) = 1-\alpha$$
$$\implies \left[\frac{(n-1)S^2}{\chi^2_{1-\alpha/2}}, \frac{(n-1)S^2}{\chi^2_{\alpha/2}}\right]$$

      Note this is not symmetric.

  - For two samples, can create CI for $\frac{\sigma_1^2}{\sigma_2^2}$. Use F_{n_1-1,n_2-1}.

- Convergence concepts
  - Convergence almost surely: $X_n \xrightarrow{a.s.} X$ if

$$P(\omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)) = 1$$

  - Convergence in probability: $X_n \xrightarrow{p} X$ if

$$\forall \epsilon > 0, P(|X_n - X| \leq \epsilon) \to 1 \text{ as } n \to \infty \iff P(|X_n - X| \geq \epsilon) \to 0$$

  - Convergence in distribution: $X_n \xrightarrow{d} X$ if

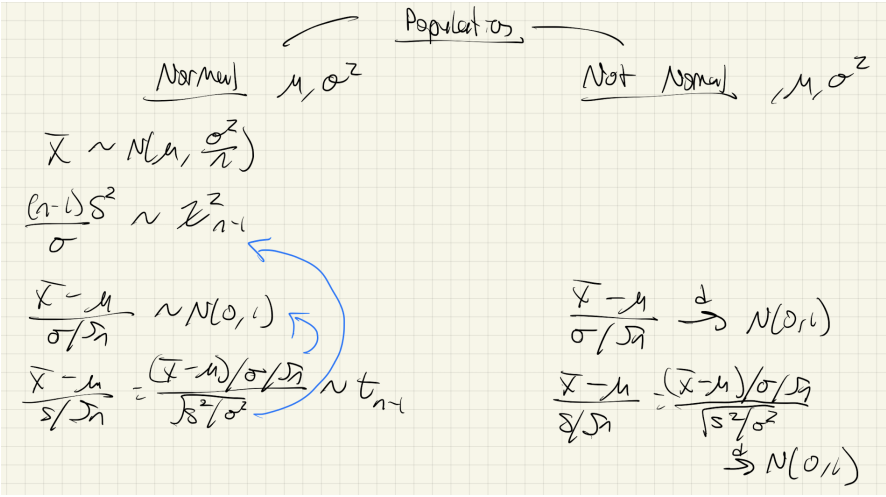
$$F_{X_n}(x) \to F_X(x) \text{ as } n \to \infty \text{ for every continuity point of } F_X$$

  - Properties and theorems
    1. If $X_n \xrightarrow{a.s.} X, \implies X_n \xrightarrow{p} X$
    2. Weak Law of Large Numbers. If $X_1, \ldots, X_n \sim iid\ F_X$ with finite mean μ and finite variance σ^2, then $\bar{X} \xrightarrow{p} \mu$. (Consistency of $\bar{X}$).
    3. If $X_n \xrightarrow{p} X, Y_n \xrightarrow{p} X$, then
       (a) $X_n + Y_n \xrightarrow{p} X + Y$
       (b) $aX_n \xrightarrow{p} aX, \quad a \in \mathbb{R}$
       (c) If $X_n \xrightarrow{p} a \in \mathbb{R}$, $g(.)$ is continuous at a, then $g(X_n) \xrightarrow{p} g(a)$ (Continuous mapping theorem)
       (d) $g(X_n) \xrightarrow{p} g(X)$ for continuous g (CMT)
       (e) $X_n Y_n \xrightarrow{p} XY$
    4. $S_n^2 \xrightarrow{p} \sigma^2$ (The sample variance from n elements is consistent)
    5. If $X_n \xrightarrow{p} X, \implies X_n \xrightarrow{d} X$
    6. If $X_n \xrightarrow{d} a, \implies X_n \xrightarrow{p} a$. One of the only times convergence in distribution implies convergence in probability.
    7. $X_n \xrightarrow{d} X, g(.)$ continuous, then $g(X_n) \xrightarrow{d} g(X)$. (CMT)
    8. If $X_n \xrightarrow{d} X, A_n \xrightarrow{p} a, B_n \xrightarrow{b}$, then $B_n X_n + A_n \xrightarrow{d} bX + a$ (Slutsky).

- Central Limit Theorem. For $X_1, \ldots X_n$ iid with mgfs existing in a neighborhood of 0, finite mean and variance μ, σ^2, if $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$, then:

$$Z_n = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} N(0,1)$$

  Prove using convergence of mgfs and Taylor expansion.
  - Construct approximate confidence intervals
  - If variance is unknown, still converges to Normal by Slutsky
  - T dist converges to Normal for large df.
  - Multivariate. $X_i \in \mathbb{R}^p,\ E(X_i) = p,\ Var(X_i) = \Sigma$, then $Z_n = \frac{\bar{X}-\mu}{1/\sqrt{n}} \xrightarrow{d} N_p(0, \Sigma)$



- Delta method. If $Y_1, Y_2, \ldots$ a sequence of random variables satisfying $\sqrt{n}(Y_n - \theta) \xrightarrow{d} N(0, \sigma^2)$, then if $g(.)$ is continuously differentiable, we have

$$\sqrt{n}\{g(Y_n) - g(\theta)\} \xrightarrow{d} N(0, [g'(\theta)]^2)\sigma^2)$$

  Prove using a Taylor expansion and O-concepts below.

  - Multivariate. If $Y_n \in \mathbb{R}^p, \sqrt{n}(Y_n - \theta) \xrightarrow{d} N_p(0, \Sigma)$, then for $g : \mathbb{R}^p \to \mathbb{R}^q, q \leq p$, then:

$$\sqrt{n}[g(Y_n) - g(\theta)] \xrightarrow{d} N_q([g'(\theta)]\Sigma[g'(\theta)]^T)$$

    for

$$g'(\theta) = \begin{bmatrix} \frac{\partial g_1}{\partial \theta_1} & \cdots & \frac{\partial g_1}{\partial \theta_p} \\ \vdots & & \vdots \\ \frac{\partial g_q}{\partial \theta_1} & \cdots & \frac{\partial g_q}{\partial \theta_p} \end{bmatrix} \in \mathbb{R}^{q \times p}$$

    Can find the asymptotic dist of S_n^2, since $S_n^2 = g(\frac{1}{n}\sum X_i^2, \bar{X})$.

- O-notation and bounded in probability
  - O-notation
    - $a_n = O(b_n) \iff \left|\frac{a_n}{b_n}\right| \leq M \iff |a_n| \leq M|b_n|\, M > 0,\ \forall n$ (big O)
    - $a_n = O(1) \iff |a_n| \leq M$
    - $a_n = o(b_n) \iff \left|\frac{a_n}{b_n}\right| \to 0$ as $n \to \infty$ (little O)
    - $a_n = o(1) \iff |a_n| \to 0$
  - Bounded in probability. A sequence of r.v.s is bounded in probability, $X_n = O_p(1)$, if $\forall \epsilon > 0,\ \exists B_\epsilon > 0$ s.t. $P(|X_n| \leq B_\epsilon) \geq 1 - \epsilon,\ \forall n \geq N_\epsilon$
    - If $X_n \xrightarrow{d} X, \implies X_n = O_p(1)$
    - If $X_n = o_p(Y_n), Y_n = O_p(1), \implies X_n \xrightarrow{p} 0$

- Order statistics. For $1 \leq j \leq n$,

$$\begin{aligned} F_{X_{(j)}}(x) &= P(X_{(j)} \leq x) \\ &\iff P(Y \geq j) \quad \text{if } Y = \{\#X_i's \leq x\} \sim Binom(n, F(x)) \\ &= \sum_{k=j}^{n} \binom{n}{k} F(x)^k (1-F(x))^{n-k} \\ f_{X_{(j)}} &= \binom{n}{j} j F(x)^{j-1}(1-F(x))^{n-j} f(x) \\ &= \frac{n!}{(j-1)!(n-j)!} F(x)^{j-1}(1-F(x))^{n-j} f(x) \end{aligned}$$

  Think about 3 bins with n observations, this is multinomial.

- Estimation. Includes moment, Bayesian, MLE.
  - Moment estimation. Match sample moments with population moments: $E(X^m) = \hat{\mu}'_m$
  - Evaluate using mean squared error: $MSE(\hat{\theta}) = E_\theta[(\hat{\theta} - \theta)^2] = Var_\theta(\hat{\theta}) + (E_\theta(\hat{\theta}) - \theta)^2 = Var_\theta(\hat{\theta}) + Bias_\theta(\hat{\theta})^2$

- Maximum Likelihood Estimation
  - Likelihood $L(\theta) = \prod_{i=1}^{n} f(X_i; \theta)$. MLE $\hat{\theta} = \arg\max_\theta L(\theta)$
  - At the true parameter value θ_0,

$$\lim_{n\to\infty} P(L(\theta_0|\mathbb{X}) > L(\theta|\mathbb{X})) = 1 \quad \forall \theta \neq \theta_0$$

    Prove this using WLLN, Jensen's inequality, and property of logs.
  - Recognize when the MLE is at the boundary of the parameter space.
  - Properties:
    * MLE is consistent: $\hat{\theta} \xrightarrow{p} \theta_0$
    * MLE is asymptotically normal: $\hat{\theta} \sim AN(\theta, \frac{1}{nI_1(\theta)}) = AN(\theta, \frac{1}{I_n(\theta)}) \iff$

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, I_1^{-1}(\theta_0))$$

    where

$$\begin{aligned} I_1(\theta) &= \text{Fisher's information (with 1 observation)} \\ &= Var\left(\frac{\partial \log f(x_1;\theta)}{\partial\theta}\right) \\ &= -E\left[\frac{\partial^2}{\partial\theta^2} \log f(x_1;\theta)\right] \\ &= \frac{1}{n} I_n(\theta) \\ &= \frac{1}{n} E\left[\left(\frac{\partial}{\partial\theta} \log L(\theta|\mathbb{X})\right)^2\right] = \frac{1}{n} E\left[U(\theta)^2\right] \end{aligned}$$

    where $U(\theta)$ is the Score function. Prove asymptotic normality using Taylor expansion of the score function evaluated at θ_0, and WLLN, regularity conditions, bounded in probability.

  - ∗ Inference: by asymptotic normality, a CI is: $\left[\hat{\theta} \pm Z_{1-\alpha/2}\frac{1}{\sqrt{nI_1(\hat{\theta})}}\right]$
  - ∗ Invariance property. If $\eta = g(\theta), \hat{\eta} = g(\hat{\theta})$, where $\hat{\theta}$ is the MLE of θ, then $\hat{\eta}$ is the MLE of η.

- Cramer-Rao lower bound. For $Y = u(\mathbb{X}), E(Y) = k(\theta)$:

$$Var(Y) \geq \frac{[k'(\theta)]^2}{nI_1(\theta)}$$

  An estimator is efficient if it obtains the C-R lower bound.

- EM Algorithm. Algorithm to find the MLE. Not guaranteed to reach it, but $L(\theta^{(m+1)}|\mathbb{X}) \geq L(\theta^{(m)}|\mathbb{X})$.

$$\begin{aligned} \text{E step:} \quad Q(\theta) &= E\left[\log L^C(\theta|\mathbb{X}, \mathbb{Z})|\hat{\theta}^{(m)}, \mathbb{X}\right] \\ \text{M step:} \quad \hat{\theta}^{(m+1)} &= \arg\max_{\theta} Q(\theta) \end{aligned}$$

- Sufficient statistics. $T(\mathbb{X})$ is sufficient for θ if the distribution of $\mathbb{X}|T(\mathbb{X})$ does not depend on θ
  - Factorization theorem. $T(\mathbb{X})$ sufficient for $\theta \iff f_{\mathbb{X}}(\mathbb{x}, \theta) = g(T(\mathbb{X}), \theta)h(\mathbb{x}) \quad \forall \mathbb{x}$
  - Rao-Blackwell. If $T(\mathbb{X})$ sufficient for θ, $u(\mathbb{X})$ unbiased for θ, then $Y = E[u(\mathbb{X})|T(\mathbb{X})]$ is unbiased with $Var(Y) \leq Var(u(\mathbb{X}))$

- Exponential family. X is in the exponential family if it has pdf of the form:

$$f(X|\theta) = h(x)c(\theta)\exp\left\{\sum_{i=1}^{k} w_i(\theta)t_i(X)\right\}$$

  - A sufficient statistic for θ is: $T(\mathbb{X}) = [\sum_{i=1}^n t_1(x_i), \ldots, \sum_{i=1}^n t_k(x_i)]$

- Bootstrapping. Have a sample $X_1, \ldots, X_n \sim F$, and are interested in characteristics/functions of F: $\theta = T(F)$. Estimate using $\hat{\theta} = T(\hat{F})$, where the empirical cdf is $\hat{F}(x) = \frac{1}{n}\sum \mathbb{1}_{X_i \leq x}$. We need a sampling distribution for $\hat{\theta}$.
  - Nonparametric bootstrap. Resample with replacement B times: for $1 \leq j \leq B$, obtain $X_1^{*(j)}, \ldots, X_n^{*(j)}, \implies T(\hat{F}^{*(j)}) = \hat{\theta}^{*(j)}$. Can estimate empirical cdf for sampling distribution from the bootstrap samples.
  - Parametric boostrap. If we are starting from a parametric model $\hat{F} = F_{\hat{\theta}}$, we can resample from $F_{\hat{\theta}}$ instead.

- Bayesian statistics. Population parameters θ are considered random variables with prior distribution $\pi(\theta)$. Draw a sample, then estiamte the posterior distribution $\pi(\theta|\mathbb{X}) = \frac{f(\mathbb{X}|\theta)\pi(\theta)}{f(\mathbb{X})}$
  - Conjugate priors are such that the posterior distribution $\theta|\mathbb{X}$ is from the same family of distributions as the sample $\mathbb{X}$.

- Generating a random variable
  - Direct method. Let $U \sim Unif(0,1)$. If $X \sim F$, then $F^{-1}(U) \sim F$.
    Possible to use for exponential, χ_p^2 for p even, gamma, beta, Normal.

- Discrete random variables. First generate $U \sim Unif(0,1)$, then let $Y = y$ if $F(y_i) < U \leq F(y_{i+1})$
- Indirect method. Use a candidate density f_V with common support. Accept/reject algorithm: To generate $Y \sim f_Y(y)$:
  1. Generate $U \sim Unif(0,1), V \sim f_V$ independently
  2. If $U < \frac{1}{M}\frac{f_Y(v)}{f_V(v)}$, then $Y = V$. Stop. Otherwise, repeat from step 1. Repeat this for each datapoint needed.

  Prove this works.

  Note:
  * $M = \frac{1}{P(stop)}, P(stop) = \frac{1}{M}$, and if X = # trials needed to generate Y, then $X \sim Geometric(\frac{1}{M})$. Therefore the optimal $M = \sup_y \frac{f_Y(y)}{f_V(y)} \leq \infty$
  * This works if we don't have the normalizing constant: $f_Y(y) = \frac{g(X)}{c}$

5.1 Hypothesis testing

- Test procedure, creating a decision rule
- Type I vs. type II error
- Power function: $\beta(\theta) = P_\theta(\text{accept } H_1)$.
  - Size of a test: $\sup_{\theta \in \Theta_0} \beta(\theta) \leq \alpha =$ max type I error rate set
  - $P(\text{type II error}) = 1 - \beta(\theta)$
- 1-sided vs. 2-sided tests
- Sample size formula. How to derive it, what are the assumptions?

$$n = \frac{(Z_\alpha + Z_{1-\beta})^2 \sigma^2}{\mu_1^2}$$

- Likelihood ratio test

$$\lambda(\mathbb{X}) = \frac{\sup_{\theta \in \Theta_0} L(\theta|\mathbb{X})}{\sup_{\theta \in \Theta} L(\theta|\mathbb{X})}$$

  Reject H_0 if $\lambda(\mathbb{X}) \leq c$

  - Wilk's theorem. Under $H_0, -2\log\lambda(\mathbb{X}) \xrightarrow{d} \chi^2_q$, where $q = dim(\Theta) - dim(\Theta_0)$. Reject H_0 if $\chi^2_{LR} > \chi^2_{1-\alpha,q}$
  - Equivalence between LRT and Z/T tests for a normal population
  - Example: derive $\lambda(\mathbb{X}, \mathbb{Y})$ for two-sample T test
  - Neyman-Pearson. For a test with rejection region $R = \{\mathbb{X} : F(\mathbb{X}|\theta_1) > kf(\mathbb{X}|\theta_0)\} = \{\mathbb{X} : \lambda(\mathbb{X}) < c\}$ and $\alpha = P_{\theta_0}(\mathbb{X} \in R)$, if R' is the rejection region for any α level test with $P_{\theta_0}(\mathbb{X} \in R') \leq \alpha$, R has larger power than R'.
- Score test. Recall the score function $U(\theta) = \frac{\partial}{\partial\theta}\log L(\theta|\mathbb{X})$. Under H_0:

$$\begin{aligned} \frac{1}{n}U(\theta_0) &\xrightarrow{d} N(0, I_1(\theta_0)) \\ \frac{U(\theta_0)}{\sqrt{I_n(\theta_0)}} &\xrightarrow{d} N(0,1) \\ X^2_S = U(\theta_0)^T I_n^{-1}(\theta_0) U(\theta_0) &\xrightarrow{d} \chi^2_p \end{aligned}$$

- Wald test. If $H_0 : A\theta = d \in \mathbb{R}^q$:

$$\begin{aligned} \chi^2_W = (A\hat{\theta} - d)^T (AI_n^{-1}(\theta)A^T)^{-1}(A\hat{\theta} - d) &\xrightarrow{d} \chi^2_q \\ \frac{W_n - \theta_0}{S_n} &\xrightarrow{d} N(0,1) \end{aligned}$$

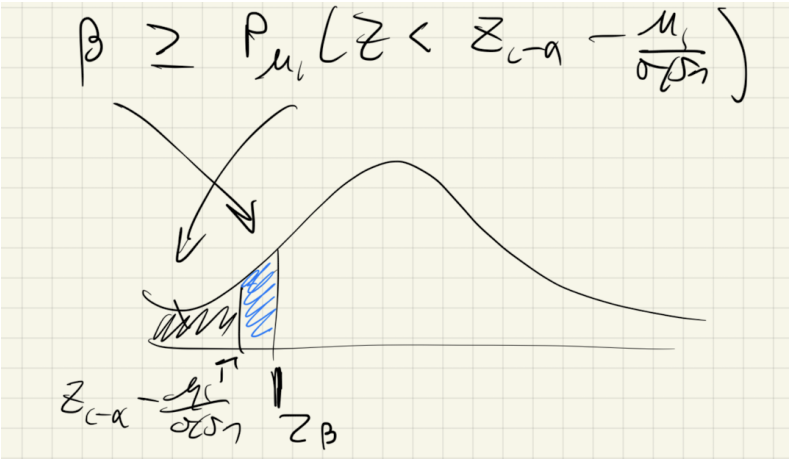
- P-values. The probability of obtaining a test statistic at least as extreme as the one observed. Prove that $p \sim Unif(0,1)$.

- Power analysis. If $H_0 : \mu = \mu_0 (= 0)$, $\quad H_1 : \mu = \mu_1 (> 0)$, start with:

$$\begin{aligned} \beta(\mu_1) &> 1 - \beta \\ \beta(\mu_0) &= \alpha \end{aligned}$$

Under normality:

$$\begin{aligned} P_{\mu_1}\left(Z > Z_{1-\alpha}\right) &\geq 1 - \beta \quad \text{(Not std normal under H1)} \\ P_{\mu_1}\left(Z > Z_{1-\alpha} - \frac{\mu_1}{\sigma/\sqrt{n}}\right) &\geq 1 - \beta \quad \text{(Expand Z and get std Normal on LHS)} \\ \beta &\geq P_{\mu_1}\left(Z \leq Z_{1-\alpha} - \frac{\mu_1}{\sigma/\sqrt{n}}\right) \\ Z_{1-\alpha} - \frac{\mu_1}{\sigma/\sqrt{n}} &\leq Z_\beta \quad \text{(Apply inverse std Normal cdf)} \end{aligned}$$



6 Other useful formulas and identities

- Geometric sum. For $|r| < 1$:

$$\begin{aligned}\sum_{k=0}^{\infty} ar^k &= \frac{a}{1-r} \\ \sum_{k=0}^{n} ar^k &= \frac{a(1-r^n)}{1-r}\end{aligned}$$

- Definitions of e

$$\begin{aligned}e^x &= \sum_{k=0}^{\infty} \frac{x^k}{x!} \\ &= \lim_{n\to\infty} \left(1+\frac{x}{n}\right)^n\end{aligned}$$

- Gamma function

$$\begin{aligned}\Gamma(\alpha) &= \int_0^{\infty} t^{\alpha-1}e^{-t}dt \\ \Gamma(\alpha+1) &= \alpha\Gamma(\alpha) \\ \Gamma(n) &= (n-1)! \\ \Gamma(1/2) &= \sqrt{\pi}\end{aligned}$$

- Beta function

$$\begin{aligned}B(\alpha,\beta) &= \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx \\ &= \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}\end{aligned}$$

- Integration by parts

$$\int udv = uv - \int vdu$$

- Taylor expansion (second order) of f around a:

$$f(x) \approx f(a) + \frac{f'(a)}{1!}(x-a) + \frac{f''(a)}{2!}(x-a)^2 + o((x-a)^2)$$

- Limit superior, limit inferior:

$$\begin{aligned}\limsup a_n &= \lim_{n\to\infty} \sup_{m\geq n}\{a_m\} \\ \liminf a_n &= \lim_{n\to\infty} \inf_{m\geq n}\{a_m\}\end{aligned}$$