0 STAT 230 review

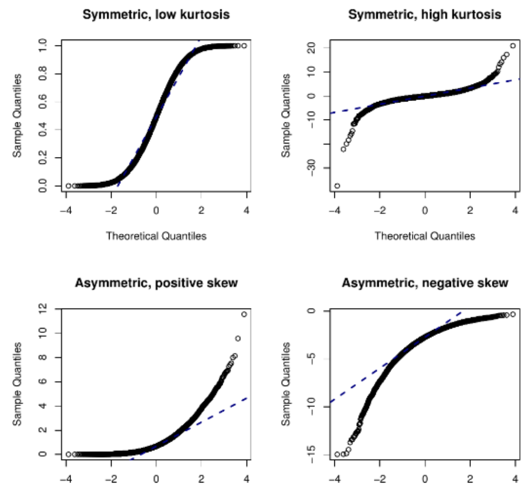
- If A, B are independent $P(A \cap B) = P(A)P(B)$.
- **Conditional Probability:** $P(A|B) = \frac{P(A\cap B)}{P(B)}$.
- **Bayes' Theorem:** $P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{P(B|A)P(A)}{P(B|\overline{A})P(\overline{A})+P(B|A)P(A)}$.
- **Variance:** $\text{Var}(X) = E[(X - E[X])^2] = E[X^2] - E[X]^2$.
- **Covariance:** $\text{cov}(X, Y) = E[(X - E[X])(Y - E[Y])]$.
- **Correlation:** $\rho_{X,Y} = \frac{\text{cov}(X,Y)}{\text{SD}(X)\,\text{SD}(Y)}$.
- $E(aX + b) = aE(X) + b$.
- $Var(aX + b) = a^2 Var(X)$.
- If X, Y are independent then $E(aX + bY) = aE(X) + bE(Y)$ and $Var(aX + bY) = a^2Var(X) + b^2Var(Y)$.
- If $N \sim G(\mu, \sigma)$ then $\texttt{pnorm}(x, \mu, \sigma) = P(N \le x)$ and $\texttt{qnorm}(x, \mu, \sigma) = a$ where $P(N \le a) = x$.
- **Central Limit Theorem:** If $X_1, \dots, X_n$ independent with $E(X_i) = \mu$ and $\text{Var}(X_i) = \sigma^2$ then $\dfrac{\sum\limits_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}} \sim G(0, 1)$ and $\dfrac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim G(0, 1)$.

1 Statistical Sciences

- **Unit:** Individual person, place, or thing we take measurements about.
- **Population:** Collection of units.
- **Process:** Ongoing system by which units are produced.
- **Variate:** Characteristics of unit that can be measured. One of discrete (countably many values), continuous (infinite precision), categorical (categories), ordinal (categories with ordering), complex (e.g., text).
- **Attribute:** Function of of a variate defined for all units.
- **Sample Survey:** Study of finite population by taking a representative sample.
- **Observational Study:** Study of population or process collected routinely over time without attempting to change any variates.
- **Experimental Study:** Study of population where specific variates are changed or fixed.
- **Sample Variance:** $s^2 = \frac{1}{n-1}\sum_{i=1}^n (y_i - \overline{y})^2 = \frac{1}{n-1}\left(\sum_{i=1}^n (y_i^2) - n\overline{y}^2\right)$
- **Sample Range:** range $= \max_i(y_i) - \min_i(y_i) = y_{(n)} - y_{(1)}$
- **Sample Quantile:** pth quantile or $100p$th percentile is the value q where $q(p) = P(X \le q) = p$.
- **Inter-Quartile Range (IQR):** IQR $= q(0.75) - q(0.25)$.
- **Sample Skewness:** $\dfrac{\frac{1}{n}\sum\limits_{i=1}^{n}(y_i - \overline{y})^3}{\left[\frac{1}{n}\sum\limits_{i=1}^{n}(y_i - \overline{y})^2\right]^{3/2}}$. Skewness < 0 implies left tail, skewness > 0 implies right tail, skewness ≈ 0 implies symmetric.
- **Sample Kurtosis:** $\dfrac{\frac{1}{n}\sum\limits_{i=1}^{n}(y_i - \overline{y})^4}{\left[\frac{1}{n}\sum\limits_{i=1}^{n}(y_i - \overline{y})^2\right]^{2}}$. Kurtosis < 3 implies light tails, kurtosis > 3 implies heavy tails, kurtosis ≈ 3 implies normal tails.
- **Normality:** Normal distribution should have: (1) mean and median approximately equal, (2) skewness ≈ 0, (3) kurtosis ≈ 3, (4) approximately 95% of observations should be in $[\overline{y} - 2s, \overline{y} + 2s]$ (5) histogram or e.c.d.f. should agree with theoretical c.d.f., (6) Q-Q plot should be approximately a straight line.
- **Five Number Summary:** $(y_{(1)}, q(0.25), q(0.5), q(0.75), y_{(n)})$.
- **Relative Risk:** $R = \frac{A_1/(A_0+A_1)}{B_1/(B_0+B_1)}$. Likelihood of presenting variate based on membership. X_0 not presenting, X_1 presenting.
- **Estimation Problem:** Estimating attributes of a population/process.
- **Hypothesis Testing Problem:** Assessing the truth of a question.
- **Prediction Problems:** Predicting future value of variate of a unit.

2 Models and MLE

- **Likelihood Function:** $L(\theta) = L(\theta; y) = P(Y = y; \theta)$ where $\theta \in \Omega$.
- **Maximum Likelihood Estimate:** $\hat{\theta}_{MLE} = \arg\max_{\theta\in\Omega} L(\theta)$.
- **Relative Likelihood Function:** $R(\theta) = \frac{L(\theta)}{L(\hat{\theta}_{MLE})}$ where $\theta \in \Omega$.
- **Log Likelihood Function:** $\ell(\theta) = \log L(\theta)$ where $\theta \in \Omega$.
- **Log Relative Likelihood Function:** $r(\theta) = \log R(\theta)$ where $\theta \in \Omega$.
- Note that $\hat{\theta}_{MLE}$ is the value that maximizes $L(\theta)$, $R(\theta)$, and $\ell(\theta)$.
- **Likelihood of Continuous Variables:** If we have i.i.d. observations $Y_1, \dots, Y_n$ then $L(\theta) = \prod\limits_{i=1}^{n} f(y_i; \theta)$ where f is the p.d.f.
- **Invariance of MLE:** $g(\hat{\theta}_{MLE})$ is the MLE of $g(\theta)$.
- **Q-Q Plot:** A plot of the points $\left(\phi^{-1}\left(\frac{i}{n+1}\right), y_{(i)}\right)$ where ϕ^{-1} is the inverse of the c.d.f. of $G(0, 1)$. If the data is approximately Gaussian this should be a straight line.



- **kth Moment:** $\mu_k = E(Y^k)$. Sample kth moment: $m_k = \frac{1}{n}\sum_{i=1}^n y_i^k$.
- **Method of Moments Estimate:** Estimate parameters by:
  1. Compute the first p sample moments where p is the number of unknown parameters.
  2. Relate the population moments to the true parameter values.
  3. Use the sample moments to solve the resulting system of equations to estimate the parameters.

3 Conducting Studies

- **PPDAC:**
  - Problem: A clear statement of the study's objectives.
  - Plan: The procedures used to carry out the study including how the data will be collected.
  - Data: The physical collection of the data, as described in the plan.
  - Analysis: The analysis of the data collected in light of the problem and the plan.
  - Conclusion: The conclusions that are drawn about the problem and their limitations.
- **Target Population:** The population (or process respectively) to which we want the conclusions to apply.
- **Study Population:** The population of units available to be included in the study. Hopefully subset of target population.
- **Study Error:** The difference in attributes between the target and study populations.
- **Sample Error:** The difference in attributes between the study population and the sample. Random samples have no sample error.
- **Measurement Error:** The difference between true values of variates and measured values of variates for units in the sample.

4 Estimation

- **Point Estimator:** Function $\tilde{\theta} = g(Y_1, \dots, Y_n)$ of observations $Y_1, \dots, Y_n$. Gives point estimate $\hat{\theta} = g(y_1, \dots, y_n)$. Distribution of $\tilde{\theta}$ is called the sampling distribution of the estimator.
- **Bias:** How much we expect an estimator to be off by. $\text{Bias}(\tilde{\theta}) = E[\tilde{\theta}] - \theta$.
- **Mean Squared Error (MSE):** Trade off between bias and variance of estimator. $\text{MSE}(\tilde{\theta}) = E[(\tilde{\theta} - \theta^2)] = \text{Var}(\tilde{\theta}) + \text{Bias}(\tilde{\theta})^2$.
- **Score Function:** $U(\theta; Y) = \frac{\partial}{\partial\theta}\ell(\theta; Y) = \frac{1}{L(\theta;Y)}\frac{\partial}{\partial\theta}L(\theta; Y)$, i.e., the slope of $\ell(\theta)$ at the true θ. $U(\theta; Y)$ is a random variable with $E[U|\theta] = 0$.
- **Fisher Information:** The variance of the score function given by $\mathcal{I}(\theta) = E\left\{\left[\frac{\partial}{\partial\theta}\ell(\theta; Y)\right]^2 \middle| \theta\right\} = -E\left[\frac{\partial^2}{\partial\theta^2}\log L(\theta; Y)\middle|\theta\right]$. Low information means blunt log-likelihood, high information means sharp log-likelihood. Low information means lots of values of $\hat{\theta}$ are similarly good to the MLE. If $Y_1, \dots, Y_n$ are i.i.d. then $\mathcal{I}(\theta) = n\mathcal{I}_1(\theta)$ where $\mathcal{I}_1(\theta) = -E\left[\frac{\partial^2}{\partial\theta^2}\log L(\theta; Y_1)\middle|\theta\right]$.
- **Cramér-Rao Lower Bound:** For any unbiased estimator $\tilde{\theta}$ of θ we have $\text{Var}(\tilde{\theta}) \ge \frac{1}{\mathcal{I}(\theta)}$.
- **Efficiency:** $e(\tilde{\theta}) = \frac{1/\mathcal{I}(\theta)}{Var(\tilde{\theta})}$ where $\tilde{\theta}$ is an unbiased estimator of θ. Note $0 < e(\tilde{\theta}) \le 1$ and if $e(\tilde{\theta}) = 1$ then $\tilde{\theta}$ is said to be efficient or be the minimum-variance unbiased estimator of θ.
- **$100p$% Likelihood Interval:** The set $\{\theta : R(\theta) \ge p\} = \{\theta : r(\theta) \ge \log p\}$. These are the values of θ which makes the data at least $100p$% as likely as if $\theta = \hat{\theta}_{MLE}$.
- **Coverage Probability:** Probability $P(\theta \in [L(Y), U(Y)]) = P(L(Y) \le \theta \le U(Y))$ where $[L(Y), U(Y)]$ is an interval estimator for θ.

- **100p% Confidence Interval:** The smallest (usually symmetric) interval estimate $[L(Y), U(Y)]$ with coverage $P(L(Y) \leq \theta \leq U(Y)) = p$.
- **Pivotal Quantity:** $Q = Q(Y; \theta)$ function of the data and unknown parameter θ such that the distribution is known.
- **Approximate Pivotal Quantity:** $Q_n = Q_n(Y_1, \dots, Y_n; \theta)$ such that as $n \to \infty$, Q_n is a known distribution (which doesn't rely on unknowns).
- **Likelihood Ratio Statistic:** $\Lambda(\theta) = -2\log\left(\frac{L(\theta;Y)}{L(\hat{\theta}_{MLE};Y)}\right) = -2r(\theta)$ is a random variable. For large n, $\Lambda(\theta) \sim \chi^2_1$ approximately.
- **Gaussian μ CI:** A 100p% confidence interval for μ given data $Y_1, \dots, Y_n \sim G(\mu, \sigma)$ is given by $\overline{y} \pm a\frac{s}{\sqrt{n}}$ where $a = \texttt{qt}(\frac{1+p}{2}, n-1)$.
- **Gaussian σ CI:** A 100p% confidence interval for σ given data $Y_1, \dots, Y_n \sim G(\mu, \sigma)$ is given by $\left[\sqrt{\frac{(n-1)s^2}{b}}, \sqrt{\frac{(n-1)s^2}{a}}\right]$ where $a = \texttt{qchisq}(\frac{1-p}{2}, n-1)$ and $b = \texttt{qchisq}(\frac{1+p}{2}, n-1)$.
- A 100q% confidence interval is approximately equivalent to a 100p% likelihood interval where $p = e^{-c/2}$ and $c = \texttt{qchisq}(q, 1) = \texttt{qnorm}(\frac{q+1}{2})^2$.
- A 100p% likelihood interval is approximately equivalent to a 100q% confidence interval where $q = \texttt{pchisq}(d, 1) = 2 \cdot \texttt{pnorm}(\sqrt{d}) - 1$ and $d = -2\log p$. Note q is the coverage of the likelihood interval.
- Important pivotal quantities:
  - $\frac{\overline{Y}-\mu}{1/\sqrt{n}} \sim G(0,1)$ (exact).
  - $\Lambda(\theta) = -2\log\left(\frac{L(\theta;Y)}{L(\hat{\theta}_{MLE};Y)}\right) \sim \chi^2_1$ (approx).
  - $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$ where $Y_1, \dots, Y_n \sim G(\mu, \sigma)$ (exact).
  - $\frac{\overline{Y}-\mu}{S/\sqrt{n}} \sim t_{n-1}$ where $Y_1, \dots, Y_n \sim G(\mu, \sigma)$ (exact).
  - $\frac{\tilde{\theta}-\theta}{\sqrt{\frac{\tilde{\theta}(1-\tilde{\theta})}{n}}} \sim G(0,1)$ where $Y \sim Bin(n, \theta)$ (approx).
- Building a 100p% confidence interval for θ:
  1. Find a pivotal quantity $Q(Y; \theta)$.
  2. Find a, b such that $P(a \leq Q(Y;\theta) \leq b) = p$. I.e. find a, b so that $P(Q \leq a) = 1 - P(Q \leq b) = \frac{1-p}{2}$.
  3. Re-express the inequality as $P(L(Y) \leq \theta \leq U(Y)) = p$.
  4. $[L(y), U(y)]$ is a 100p% confidence interval for θ given the observed data y.

5 Hypothesis Testing

- **Null Hypothesis:** Default hypothesis; H_0.
- **Alternate Hypothesis:** Hypothesis to be tested; H_A.
- **Discrepancy Measure:** Function of the data $D = g(Y)$ which measures agreement between data and H_0. $d = g(y) \approx 0$: high agreement with H_0; $d >> 0$: high disagreement. Commonly $D = |Y - E[Y]|$.
- **p-value:** The value $P(D \geq d; H_0)$. I.e., the probability that data assuming H_0 are at least as surprising as our observed data. If $p \approx 0$ we are surprised if H_0 is true.
- $p-$value $\geq 1 - q$ iff θ_0 is in the 100q% confidence interval of θ.
- **Type 1 error:** Rejecting H_0 when H_0 is actually true. (False positive rejection.) Type 1 error rate is $P(p < \alpha) = \alpha$.
- **Type 2 error:** Accepting H_0 when H_0 is actually false. (False negative rejection.) Type 2 error rate is denoted β.
- **Power:** Probability to reject H_0 when H_0 is actually false (True positive rejection.): power $= 1-\beta$. This is the ability to recognize a signal (weird data).
- Testing a Hypothesis (general):
  1. Specify the null hypothesis H_0 and propose a model.
  2. Specify a discrepancy measure $D(Y)$ where $D >> 0$ corresponds to data inconsistent with H_0. Compute $d = D(y)$.
  3. Calculate $p-$value $= P(D \geq d; H_0)$.
  4. Draw conclusions.
- Testing $H_0: \mu = \mu_0$ given data $Y_1, \dots, Y_n \sim G(\mu, \sigma)$.
  1. Use $D = \frac{|\overline{Y}-\mu_0|}{S/\sqrt{n}}$ to compute $d = D(y)$.
  2. Calculate $p-$value $= 2[1 - P(T \leq d)]$ where $T \sim t_{n-1}$.
  3. Draw conclusions.
- Testing $H_0: \sigma = \sigma_0$ given data $Y_1, \dots, Y_n \sim G(\mu, \sigma)$.
  1. Use $U = \frac{(n-1)s^2}{\sigma_0^2}$ to compute $u = U(y)$.
  2. Compute $P(U \leq u)$ for $U \sim \chi^2_{n-1}$.

  3a. If $P(U \leq u) < 0.5$ then $p-$value $= 2P(U \leq u)$.
  3b. If $P(U \leq u) < 0.5$ then $p-$value $= 2(1 - P(U \leq u))$.
- Testing $H_0: \theta = \theta_0$ using likelihood ratio statistic:
  1. Find $L(\theta)$ and the MLE $\hat{\theta}$.
  2. Compute $\lambda(\theta_0) = -2\log(\frac{L(\theta_0)}{L(\hat{\theta})})$.
  3. Then p-value $= 1 - P(W \leq \lambda(\theta_0))$ for $W \sim \chi^2_1$ approximately.

6 Gaussian Response Models

- **Residual:** The vertical distance between a point and a fitted line.
- $S_{xy} = \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = n\,\mathrm{Cov}(x, y)$. In particular $SS = S_{yy}$.
- **Least-Square Estimate:** The predictions $\hat{y_i} = \mu(x_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}$ which minimizes the sum of square residuals. Assumes $Y_i \sim G(\mu(x_i), \sigma)$ (homoscedasticity). Where $k = 1$, we have $Y \sim G(\alpha + \beta x, \sigma)$ and $\hat{\beta} = \frac{S_{xy}}{S_{xx}}$ and $\hat{\alpha} = \overline{y} - \hat{\beta}\overline{x}$, these are also the MLEs. $\hat{\beta_j}$ represents the increase in the mean of the response variate for a one unit increase in the explanatory variate x_j when the other variates are fixed.
- **Sum of Square Errors/Residuals:** $SSE = \sum_{i=1}^{n}(y_i - \hat{y_i})^2$.
- **Sum of Square Regressions:** $SSR = \sum_{i=1}^{n}(\hat{y_i} - \overline{y})^2 = S_{yy} - SSE$.
- **Mean Squared Error** $s_e^2 = \frac{1}{n-k-1}\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{k}\beta_j x_{ij}\right)^2$. If $k = 1$ then $s_e^2 = \frac{1}{n-k-1}(S_{yy} - \hat{\beta}S_{xy})$. Note: $E[S_e^2] = \sigma^2$ and $\frac{(n-2)S_e^2}{\sigma^2} \sim \chi^2_{n-2}$.
- $\tilde{\beta}$ **Distribution**: If $Y_i \sim G(\alpha + \beta x_i, \sigma)$, then $\tilde{\beta} = \frac{1}{S_{xx}}\sum_{i=1}^{n}(x_i - \overline{x})Y_i$ So $\tilde{\beta} \sim G(\beta, \frac{\sigma}{\sqrt{S_{xx}}})$ and $\frac{\tilde{\beta}-\beta}{S_e/\sqrt{S_{xx}}} \sim t_{n-2}$.
- **Simple Linear Regression Tests and Intervals:**
  - $H_0: \beta = \beta_0$ has p-value $= 2\left[1 - P\left(T \leq \frac{|\hat{\beta}-\beta_0|}{s_e/\sqrt{S_{xx}}}\right)\right]$ for $T \sim t_{n-2}$.
  - β has a (100p%) CI of $\hat{\beta} \pm a\frac{s_e}{\sqrt{S_{xx}}}$ for $a = \texttt{qt}(\frac{1+p}{2}, n-2)$.
  - α has a CI of $\hat{\alpha} \pm a s_e\sqrt{\frac{1}{n} + \frac{(\overline{x})^2}{S_{xx}}}$ for $a = \texttt{qt}(\frac{1+p}{2}, n-2)$.
  - $\mu(x)$ has a CI of $\hat{\mu}(x) \pm a s_e\sqrt{\frac{1}{n} + \frac{(x-\overline{x})^2}{S_{xx}}}$ for $a = \texttt{qt}\frac{1+p}{2}, n-2)$.
  - $\mu(x)$ has a PI of $\hat{\mu}(x) \pm a s_e\sqrt{1 + \frac{1}{n} + \frac{(x-\overline{x})^2}{S_{xx}}}$ for $a = \texttt{qt}\frac{1+p}{2}, n-2)$.
- R^2 **Statistic:** $R^2 = 1 - \frac{SSE}{S_{yy}} = \frac{SSR}{SS} = \frac{\text{Variation explained by regression}}{\text{Total variation}}$. We want R^2 close to 1, as this explains more variation.
- **Adjusted** R^2: Adjusted $R^2 = 1 - \frac{SSE/(n-k-1)}{S_{yy}/(n-1)}$. Compensates for the fact that adding more variables can artificially improve R^2.
- **Model Checking:** Need Y_i to have Gaussian distribution with constant variance (homoscedasticity) and $E[Y_i] = \mu(x_i)$ to be linear in x_i. Can check using graphics: should be linear and evenly spread. Using residual plots: residual $\hat{r_i} = y_i - \hat{y_i}$ should be drawn from $G(0, \sigma)$ and standardized residual $\hat{r_i^*} = \frac{\hat{r_i}}{s_e}$ should be drawn from $G(0, 1)$. Plotting standardized residual plot $(x_i, \hat{r_i^*})$, 99.7% of points should be in $(-3, 3)$ and should be evenly spread out around 0. Plot $(\hat{\mu_i}(x_i), \hat{r_i^*})$ for multiple linear regression. Can also check normality of residuals using Q-Q plots.
- **Regression Pitfalls:** (1) Multicolinearity: when 2 or more variates are highly correlated, can lead to incorrect conclusions. (2) Predicting beyond covariate range: model assumption may not hold, lack of data.
- **Generalized Linear Model (GLM):** (1) Probability distribution for response variate, (2) linear model $\eta = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$, (3) link function from linear model to parameters of outcome distribution.
- **Odds:** The odds of event A are $\text{odds}(A) = \frac{P(A)}{1-P(A)}$. Prefer odds at times since $\text{odds}(A) \in \mathbb{R}$, not just $[0, 1]$.
- **Logit** $g: [0, 1] \to \mathbb{R}$ with $g(p) = \text{logit}(p) = \log(\text{odds}(p)) = \log(\frac{p}{1-p})$, also called log odds. Has inverse $g^{-1}(x) = \frac{1}{1+e^{-x}}$.
- **Log Odds Ratio:** If O_1, O_2 are the odds of events, then the log odds ratio $\log(\frac{O_1}{O_2})$ is positive if event 1 is more probable than event 2.
- **Logistic Regression:** GLM with logit as the link function. If we consider outcomes as $Y_i \sim Bin(1, p_i)$, we can fit $\text{logit}(p_i) = \eta_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$ and then recover $p_i = \frac{1}{1+e^{-\eta_i}}$. We can interpret $\hat{\beta_j}$ as the increase in log odds or equivalently as the log odds ratio. So $\hat{\beta_j} > 0$ if and only if p_i increases as x_{ij} increases. Assumes events are independent (i.e., $Y_1, \dots, Y_n$), linear regression if appropriate for the log odds.
- **Logistic Regression Model Checking:** Split the events y_i by quantiles along a covariate with $p = \frac{\text{successes}}{\text{events}}$ for each quantile. Plot the log odds against the median of the covariate in each quantile. If the relationship is linear, logistic regression seems appropriate.
- **Two Sample Gaussian (Equal)** Testing $H_0: \mu_1 - \mu_2 = 0$ given $Y_{1,1}, \dots, Y_{1n_1} \sim G(\mu_1, \sigma)$ and independently $Y_{2,1}, \dots, Y_{2n_2} \sim G(\mu_2, \sigma)$.
  1. Note $\tilde{\mu_1} - \tilde{\mu_2} = \overline{Y_1} - \overline{Y_2} \sim G(\mu_1 - \mu_2, \sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}})$
  2. Compute $S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}$ so that $E[S_p^2] = \sigma^2$.
  3. Then $\frac{\overline{Y_1} - \overline{Y_2} - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$.

  4a. CI for $\mu_1 - \mu_2$ is $\overline{y_1} - \overline{y_2} \pm a s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ for $a = \texttt{qt}(\frac{1+p}{2}, n_1 + n_2 - 2)$.

4b. p-value $= 2[1 - P(T \le d)]$ for $d = \frac{|\overline{y_1} - \overline{y_2} - 0|}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ and $T \sim t_{n_1+n_2-2}$.

- **Two Sample Gaussian (Unequal)** Testing $H_0 : \mu_1 - \mu_2 = 0$ given $Y_{1,1}, \ldots, Y_{1n_1} \sim G(\mu_1, \sigma_1)$ and independently $Y_{2,1}, \ldots, Y_{2n_2} \sim G(\mu_2, \sigma_2)$.
  1. $\frac{\overline{Y_1} - \overline{Y_2} - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim G(0, 1)$ approximately for $n_1, n_2 \gtrapprox 30$.

  2a. CI for $\mu_1 - \mu_2$ is $\overline{y_1} - \overline{y_2} \pm a\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ for $a = \texttt{qnorm}(p)$.

  2b. p-value $= 2[1 - P(Z \le d)]$ for $d = \frac{|\overline{y_1} - \overline{y_2} - 0|}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$ and $Z \sim G(0, 1)$.

- **Two Sample Gaussian (Paired)** Testing $H_0 : \mu_1 - \mu_2 = 0$ given $Y_{1,1}, \ldots, Y_{1n_1} \sim G(\mu_1, \sigma)$ and $Y_{2,1}, \ldots, Y_{2n_2} \sim G(\mu_2, \sigma)$.
  1. Set $Y_i = Y_{1i} - Y_{2i} \sim G(\mu_1 - \mu_2, \sigma)$. Check $y_1, \ldots, y_n$ is Gaussian.

  2a. CI for $\mu = \mu_1 - \mu_2$ is $\overline{y} \pm a\frac{s}{\sqrt{n}}$ for $a = \texttt{qt}(\frac{1+p}{2}, n-1)$.

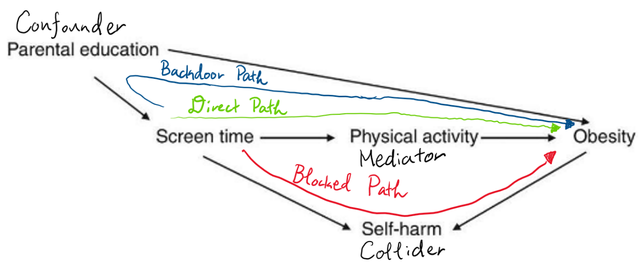
  2b. p-value $= 2[1 - P(T \le d)]$ for $d = \frac{|\overline{y} - 0|}{s/\sqrt{n}}$ and $T \sim t_{n-1}$.

7 Multinomial Models and Goodness of Fit Tests

- **Multinomial MLE:** Multinomial has $L(\theta) \propto \theta_1^{y_1} \cdots \theta_k^{y_k}$ and $\hat{\theta}_j = \frac{y_j}{n}$.
- **Pearson's Goodness of Fit Statistic:** $D = \sum_{j=1}^k \frac{(Y_j - E_j)^2}{E_j}$.
- **Degrees of Freedom:** Number of values which are free to move.
- Testing $H_0 : \theta_j = \frac{E_j}{n}$ where $e_j \ge 5$:
  1. Compute either $\lambda = 2\sum_{j=1}^{k} y_j \log(\frac{y_j}{e_j})$ or $d = \sum_{j=1}^k \frac{(y_j - e_j)^2}{e_j}$.
  2. p-value $\approx 1 - P(W \le \lambda) \approx P(W \le d)$ for $W \sim \chi^2_{k-1-p}$ where $k =$ # of categories and $p =$ # of estimated parameters.
- Testing independence in two-way table: assume categories are $A_1, \ldots, A_a$ and $B_1, \ldots, B_b$.
  1. Let r_i be the sum of row i, c_j be the sum of column j.
  2. Let $\alpha_i = P(A_i)$ and $\beta_j = P(B_j)$ with MLE $\hat{\alpha_i} = \frac{r_i}{n}$ and $\hat{\beta_j} = \frac{c_j}{n}$.
  3. $Y_{11}, Y_{12}, \ldots, Y_{ab} \sim \mathit{Multinomial}(n; \theta_{11}, \theta_{12}, \ldots, \theta_{ab})$. Independence iff $H_0 : \theta_{ij} = \alpha_i\beta_j$ is true.
  4. Expected count for $A_i \cap B_j$ is $E_{ij} = n\alpha_i\beta_j$ so $e_{ij} = \frac{r_i c_j}{n}$.
  5. Compute $\lambda = 2\sum_{i=1}^{a}\sum_{j=1}^{b} y_{ij}\log(\frac{y_{ij}}{e_{ij}})$.
  6. If $e_{ij} \ge 5$, then p-value $\approx 1 - P(W \le \lambda)$ for $W \sim \chi^2_{(a-1)(b-1)}$.

8 Causality

- **Possible Relations Between Variates:**
  1. Explanatory variate is the direct cause of the response variate.
  2. Response variate is the direct cause of the explanatory variate.
  3. Explanatory variate is a contributing cause of the response variate.
  4. Both variates are changing with time.
  5. The association is due to coincidence.
  6. Both variates have a common cause.
- **Confounding Variate:** When two variates have a common cause, the cause is called a confounding variate or confounder.
- **Dealing with Confounders:**
  - Twin studies: place one identical twin in each group.
  - Matching: find similar units from each group.
  - Randomization: randomly associate each unit with a group. This could lead to disproportionate groups, or be unethical.
- **Establishing Causation in Observational Studies:**
  1. The association between variates must be observed in many studies of different types among different groups.
  2. The association must hold when the effects of plausible confounders are taken into account.
  3. There must be a plausible scientific explanation for the direct influence of one variate on the other.
  4. There must be a consistent response.
- **Counterfactual:** The effect that would have happened in the other case. E.g., $Y(0)$ for didn't take drug, $Y(1)$ for did take drug.
- **Average Causal Effect:** $\tau = E[Y(1) - Y(0)] = \frac{1}{n}\sum_{i=1}^{n}(Y(1) - Y(0))$.
- **Propensity Score:** $\pi(x) = P(A = 1|X = x)$ where A is group, X is a variate. Often we estimate this by logistic regression.
- **Inverse Probability Weighting (IPW):** $\hat{\mu}_1 = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i \cdot 1_{(A_i=1)}}{\hat{\pi}(x_i)}$ and $\hat{\mu}_1 = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i \cdot 1_{(A_i=0)}}{1-\hat{\pi}(x_i)}$ are such that $E[\tilde{\mu_0}] = E[Y(0)]$ and $E[\tilde{\mu_1}] = E[Y(1)]$. Assumptions:
  1. Consistency: $Y_i = Y_i(0)1_{(A_i=0)} + Y_i(1)1_{(A_i=1)}$.
  2. Stable Unit Treatment Value Assumption (SUTVA): one patient receiving a treatment doesn't affect other patients' treatment.
  3. No Unmeasured Confounder (NUC): every confounding variate is accounted for in the model.
  4. Positivity: $0 < \pi(x) < 1$ for all x. Every subject has a non-zero chance of assignment to each treatment.
- **DAGs**



- Can close an open (backdoor or direct) path by accounting for it in the model. Should only do this to backdoor paths. Can open a closed (blocked) path by accounting for it in the model. Should not do this.

Distributions

- Gaussian: (Continuous) Has p.d.f. $\frac{1}{\sqrt{2\pi}\sigma}e^{-(x-\mu)^2/(2\sigma^2)}$. Arises from central limit theorem.
- χ^2_k: (Continuous) $k \geq 1$ denotes the degrees of freedom. Has p.d.f. $f(x;k) = \frac{1}{2^{k/2}\Gamma(k/2)}x^{(k/2)-1}e^{-x/2}$. Properties:
  1. If $W_1, \dots, W_n$ are i.i.d. with $W_i \sim \chi^2_{k_i}$, then $\sum_{i=1}^{n} W_i \sim \chi^2_{\sum k_i}$.

  2a. If $Z \sim G(0,1)$, then $Z^2 \sim \chi^2_1$. So, $P(W \geq w) = 2(1 - P(z \leq \sqrt{w}))$ and $P(W \leq w) = 2P(Z \leq \sqrt{w}) - 1$.

  2b. If $Z_1, \dots, Z_n \sim G(0,1)$, then $\sum_{i=1}^{n} Z_i^2 \sim \chi^2_n$.

  3. If $W \sim \chi^2_2$, then $W \sim Exp(2)$.
- Student t_k: (Continuous) $k \geq 1$ denotes the degrees of freedom. Has p.d.f. $f(x;k) = \frac{\Gamma(\frac{k+1}{2})}{\sqrt{k\pi}\Gamma(\frac{k}{2})}\left(1 + \frac{t^2}{k}\right)^{-(k+1)/2}$. Properties:
  1. $\lim_{k\to\infty} t_k \sim G(0,1)$.
  2. If $Z \sim G(0,1)$ and $U \sim \chi^2_k$, then $\frac{Z}{\sqrt{U/k}} \sim t_k$.
- Exponential: (Continuous) Assuming Poisson process for events which occur on average θ times per time unit, X denotes the number of time units before the first event occurs. Has c.d.f. $1 - e^{-x/\theta}$.
- Poisson: (Discrete) Number of events which take place in a given period of time, where on average θ events take place. X denotes the number of events. Often we assume a Poisson process:
  1. Independence: Events are independent from each other.
  2. Individuality: As the time frame Δt goes to zero, the number of events goes to zero.
  3. Uniformity: Events occur at a uniform rate over time.
- Binomial: (Discrete) Performing n Bernoulli (success/failure) trials, each with a p chance of success. X denotes the number of successes.
- Bernoulli: (Discrete) Binomial with $n = 1$.
- Negative Binomial: (Discrete) Performing Bernoulli (success/failure) trials, each with a p chance of success, until we get k successes. X denotes the number of failures before getting k successes.
- Geometric: (Discrete) Negative Binomial with $k = 1$.
- Hypergeometric: (Discrete) Drawing n objects (without replacement) from a group of N total objects, r of which are considered a success. X denotes the number of drawn successes.
- Multinomial: (Discrete) Performing n trials with k outcomes, each outcome having probability p_i. X_i denotes the number of events of type i. $f(x_1, \dots, x_k) = \frac{n!}{x_1!\cdots x_k!}p_1^{x_1}\cdots p_k^{x_k}$; $E[X_i] = np_i$; $\text{Var}[X_i] = np_i(1-p_i)$.
- Uniform: (Continuous) Drawing randomly and uniformly from an interval. X denotes the drawn value. Has c.d.f. $\frac{x-a}{b-a}$.

Distribution	p.d.f.	$E[X]$	$\text{Var}[X]$
Gaussian(μ,σ)	See above	μ	σ^2
χ^2_k	See above	k	$2k$
t_k	See above	0 if $k \geq 2$	$\frac{k}{k-2}$ if $k \geq 3$
Exponential(θ)	$\frac{1}{\theta}e^{-x/\theta}$	θ	θ^2
Poisson(θ)	$\frac{e^{-\theta}\theta^x}{x!}$	θ	θ
Binomial(n,p)	$\binom{n}{x}p^x(1-p)^{n-x}$	np	$np(1-p)$
Bernoulli(p)	$p^x(1-p)^{1-x}$	p	$p(1-p)$
NegativeBinomial(k,p)	$\binom{x+k-1}{x}p^k(1-p)^x$	$\frac{k(1-p)}{p}$	$\frac{k(1-p)}{p^2}$
Geometric(p)	$p(1-p)^x$	$\frac{1-p}{p}$	$\frac{1-p}{p^2}$
Hypergeometric(N,r,n)	$\frac{\binom{r}{x}\binom{N-r}{n-x}}{\binom{N}{n}}$	$\frac{nr}{N}$	$\frac{nr}{n}(1-\frac{r}{N})\frac{N-n}{N-1}$
Uniform(a,b)	$\frac{1}{b-a}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$

R Commands

Distribution	p.d.f.	c.d.f.	Quantiles
Z	$P(Z=z)$ or $f(z)$	$P(Z \leq z)$	$P(Z \leq a) = z$
$G(\mu,\sigma)$	`dnorm`(z,μ,σ)	`pnorm`(z,μ,σ)	`qnorm`(z,μ,σ)
χ^2_k	`dchisq`(z,k)	`pchisq`(z,k)	`qchisq`(z,k)
t_k	`dt`(z,k)	`pt`(z,k)	`qt`(z,k)
Exponential(θ)	`dexp`(z,θ)	`pexp`(z,θ)	`qexp`(z,θ)
Poisson(θ)	`dpois`(z,θ)	`ppois`(z,θ)	`qpois`(z,θ)
Binomial(n,θ)	`dbinom`(z,n,θ)	`pbinom`(z,n,θ)	`qbinom`$(z,n\theta)$

Other suffixes include `hyper` for hypergeometric, `geom` for geometric, `nbinom` for negative binomial, `unif` for uniform (continuous).

```
> mod <- lm(y ~ x1 + x2)
> summary(mod)
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.01375    5.01527  -0.202  0.84133
x1           0.73142    0.07664   9.544 3.83e-10 ***
x2           0.28225    0.09850   2.866  0.00797 **
---

Residual standard error: 4.608 on 27 degrees of freedom
Multiple R-squared:  0.9244,    Adjusted R-squared:  0.9188
```

- `Estimate` is $\hat{\beta}$ for given covariate ($\hat{\alpha}$ for intercept).
- `Std. Error` is $SD(\tilde{\beta}_i) = \frac{s_e}{\sqrt{S_{x_ix_i}}}$ for given i.
- `t value` is test statistic $t = \frac{\texttt{Estimate}}{\texttt{Std. Errror}}$.
- `Pr(>|t|)` is $2 * (1 - $ `pt(|t value|, df)`$)$ is p-value for $H_0 : \beta_i = 0$.
- `Residual standard error` is s_e.
- `Multiple R-squared` and `Adjusted R-squared` is R^2 and adjusted R^2.

`predict(mod, data.frame(x1=`x_1`,...,xn=`x_n`), interval=type, level=`p`)` where `type` $\in$ {`"confidence"`,`"prediction"`} and $\alpha \in (0,1)$ gives $100p\%$ `type` interval for covariates $x_1, \dots, x_n$.