Announcements

- The cumulative final will be in Science Center Hall B on Monday, May 8, from 2:00 pm to 5:00 pm.
- Four double-sided reference sheets are allowed.
- 4 previous finals and solutions are on Canvas.
- The previous midterms are also good reviews.
- If you've exhausted all else, Will's section notes are entirely practice problems and are (I think) mutually exclusive with other sections' problems.

Section 1: Models, metrics, likelihood

Terminology

Estimand: The thing to infer (usually a parameter in a model).

- Example: Average height of an American male.

Estimator: A function of the random variable(s) to to estimate the estimand.

- Example: A plan to take the mean height of a sample of American males.

Estimate: The crystallized estimator from observed data.

- Example: The observed sample mean (e.g. 177 cm).

Statistic: A function of the uncrystallized random variables (and possibly other constants).

- A statistic cannot involve the parameters, but its distribution can and usually should involve the parameters.

Parametric: A model in which the (joint) CDF is fully known once a finite set of parameters is specified.

- Example: $\mathcal{N}(\mu, \sigma^2)$.

Nonparametric: A model in which the (joint) CDF is only known once an infinite set of parameters is specified.

- Example: The family of all CDFs.

Estimator metrics

Bias: $\text{Bias}(\hat{\theta}) = E(\hat{\theta}) - \theta$.

- In words: The average distance of an estimator from the estimand (lower absolute bias is better).
- Unbiased: $\hat{\theta}$ is unbiased if $\text{Bias}(\hat{\theta}) = 0$.
- Strategy: Apply linearity of expectation to the estimator.

Standard error: $\text{SE}(\hat{\theta}) = \sqrt{\text{Var}(\hat{\theta})}$.

- In words: A measure of how variable the estimator is (smaller is better).
- Strategy: Solve for $\text{Var}(\hat{\theta})$ using the usual strategies (pulling out constants squared, Eve's law, etc.) and square root the result.

Mean squared error: $\text{MSE}(\hat{\theta}) = E\left((\hat{\theta} - \theta)^2\right) = \text{Var}(\hat{\theta}) + \left(\text{Bias}(\hat{\theta})\right)^2$.

- Association: Overall measure of how good an estimator is.

- RMSE: $\sqrt{\text{MSE}(\hat{\theta})}$.
- Strategy: Usually, finding the bias and standard error and then converting to MSE works best.

Consistency: $\hat{\theta} \xrightarrow{p} \theta$ as $n \to \infty$.

- In words: As we obtain more and more data, our estimator estimates the estimand better and better.
- Strategy: Show $\text{MSE}(\hat{\theta}) \to 0$.
- Strategy: Apply the Law of Large Numbers (especially useful for consistency of estimators that can be interpreted as sample means).
- Example: If $Y_1, ..., Y_n \sim \mathcal{N}(0, \sigma^2)$, $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n Y_i^2$ is consistent for σ^2 since $\frac{1}{n}\sum_{i=1}^n Y_i^2 \to E(Y_i^2) = \sigma^2$.
- Strategy: Apply the Continuous Mapping Theorem if a you want to show a function of a simple estimator converges to a function of a simple estimand.
- Example: With Y_i as above, $\hat{\sigma}$ is consistent for σ by the CMT with $g(\sigma^2) = \sqrt{\sigma^2}$.
- Strategy (Worst case): Directly verify $P(|\hat{\theta} - \theta| > \epsilon) \to 0$ as $n \to \infty$.

Likelihood

Likelihood function: $L(\theta; y) = f_Y(y|\theta)$.

- In words: The likelihood of observing the given data for a particular set of parameters as a function of the parameters, treating the data as fixed.
- Note: f_Y can be a PDF, PMF, or a joint PDF or PMF.
- Note: We usually remove *multiplicative* constants that do not involve the parameters (i.e. only involve the data or numerical constants).
- Example: For the Y_i above, $L(\sigma^2; \vec{y}) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{y_i^2}{\sigma^2}\right)$.
- Equivalent: $L(\sigma^2; \vec{y}) = \prod_{i=1}^n \frac{1}{\sqrt{\sigma^2}} \exp\left(-\frac{1}{2}\frac{y_i^2}{\sigma^2}\right)$.
- Not equivalent: $L(\sigma^2; \vec{y}) = \prod_{i=1}^n \exp\left(-\frac{1}{2}\frac{y_i^2}{\sigma^2}\right)$ (σ^2 multiplicative term is dropped).
- Not equivalent: $L(\sigma^2; \vec{y}) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{y_i^2}{\sigma^2}\right)$ (the dropped 1/2 term is not multiplicative).

Log-likelihood: $\ell(\theta; y) = \log(L(\theta; y))$.

- Strategy: Log is an order-preserving transformation, so maximizing the log likelihood is the same as maximizing the likelihood and is usually easier. (Note that log is always the natural logarithm in this class.)
- Note: We usually remove *additive* constants that do not involve the parameters.

Reparameterization: For $\psi = g(\theta)$, $L(\psi; y) = L(\theta; y)$.

- Note: The likelihood functions will not actually look the same, but evaluating $L(\theta; y)$ at a particular θ will give the same value as $L(\psi; y)$ evaluated at the corresponding ψ.
- Strategy: Plug in $g^{-1}(\psi)$ for θ everywhere in $L(\theta; y)$.
- Example: $L(\sigma^2; \vec{y}) = L(\sigma; \vec{y})$.

Data transformation: If the original $\vec{y}$ can be reconstructed from $h(\vec{y})$, $L(\theta; y) = L(\theta; h(y))$.

- Example: $L(\sigma^2; \vec{y}) = L(\sigma^2; \overrightarrow{\exp(y)})$.

Empirical CDF: $\hat{F}(y) = \frac{1}{n}\sum_{j=1}^n I(y_j \leq y)$.

Censored data: Some of the data is observed and some is in an unobserved region.

- Strategy: Find the likelihood function by multiplying the PDF/PMF of known values by the CDF (or difference in CDFs) for the unobserved region.
- Let $Y_1, ..., Y_n \sim \text{Expo}(\lambda)$ be observed particle emission times from radioactive decay, but suppose the detector's clock broke from time t_1 to t_2 so we only know that m decays occurred during that time, not when they were. Then,

$$L(\lambda; \vec{y}) = \left(\prod_{i=1}^n f_Y(y_i)\right)(F_Y(t_2) - F_Y(t_1))^m = \left(\prod_{i=1}^n \lambda e^{-\lambda y_i}\right)\left(1 - e^{-\lambda t_2} - (1 - e^{-\lambda t_1})\right)^m$$

Section 2: MLE and MOM

Maximum likelihood estimation

Maximum likelihood estimate: The value of $\hat{\theta}$ that maximizes $L(\theta; y)$.

- Note: $\hat{\theta}$ can refer to either the estimator or the estimate.
- Note: Under regularity conditions (mainly, the support of Y does not depend on θ and all expectations and derivatives of the likelihood function exist), the MLE is consistent, asymptotically Normal, asymptotically unbiased, and asymptotically efficient (no other asymptotically unbiased estimator will have a lower standard error asymptotically).
- Strategy: Write the likelihood function, convert it to the log likelihood function, set its derivative with respect to θ to 0, and solve for $\hat{\theta}$.
- Example: For $Y_1, ..., Y_n \sim \mathcal{N}(\mu, \sigma^2)$, $\hat{\mu} = \frac{1}{n}\sum_{i=1}^n Y_i$ and $\hat{\sigma^2} = \frac{1}{n}\sum_{i=1}^n (Y_i - \bar{Y})^2$.

Invariance of the MLE: If $\hat{\theta}$ is the MLE of θ, $g(\hat{\theta})$ is the MLE of $g(\theta)$.

- Note: If g is one-to-one, this follows from invariance of likelihood; otherwise, we define this to be true.
- Strategy: Use this to find non-standard quantities for a distribution (e.g. the probability a $\text{Pois}(\lambda)$ random variable is 0 using $\hat{\lambda}$ or the 0.95 quantile of $\mathcal{N}(\mu, \sigma^2)$ using $\hat{\mu}$ and $\hat{\sigma}^2$).

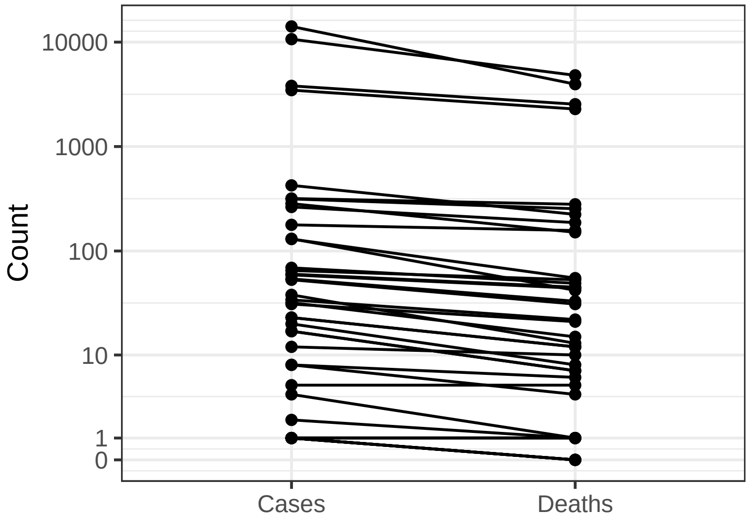
Method of moments

Method of moments: Match the sample moments with the theoretical moments.

- Strategy: Write the parameter of interest in terms of the distribution's moments and replace the true moments with sample moments.
- Example: For $Y_1, ..., Y_n \sim \mathcal{N}(\mu, \sigma^2)$, $\hat{\mu} = \frac{1}{n}\sum_{i=1}^n Y_i$ and $\hat{\sigma^2} = \left(\frac{1}{n}\sum_{i=1}^n Y_i^2\right) - \left(\frac{1}{n}\sum_{i=1}^n Y_i\right)^2$.

Problem 1

Ebola virus disease is a severe illness with a fatality rate of around 50%. The virus was first discovered in 1976, and the 2014-2016 West Africa outbreak was the most severe outbreak in recorded history. The World Health Organization tracks Ebola outbreaks by country and by year. The plot below shows the case and death counts per year and per country for years in which a country had at least one confirmed case.



It is important for government policymakers to be able to predict how many people will die of Ebola in a particular country during a particular year. Consider the following two models where Y_i is the number of Ebola deaths and N_i is the number of Ebola cases in a particular country in a particular year:

$$\begin{aligned} &\text{Model 1} : Y_i \sim \text{FS}(\lambda_1) \\ &\text{Model 2} : Y_i \sim \text{Bin}(N_i, p),\ N_i \sim \text{FS}(\lambda_2) \end{aligned}$$

Let $\mu = E(Y_1)$, and suppose we observe i.i.d. pairs $(Y_1, N_1), ..., (Y_n, N_n)$.

1. Find μ in terms of the model parameters for each model.

In Model 1, $\mu = \frac{1}{\lambda_1}$ by the mean of a first success. In model 2, by Adam's law,

$$\mu = E(E(Y_1|N_1)) = E(N_1 p) = \frac{p}{\lambda_2}$$

2. (Skip) Show that the variances of the marginal distributions of Y are not the same and use this to conclude that the marginal distribution of Y is different between the two models.

In the first model, $\text{Var}(Y) = (1-\lambda_1)/\lambda_1^2$. If the marginal distributions were equal, we could plug in $\frac{\lambda_2}{p}$ for λ_1 to get

$$\text{Var}(Y) = \left(1 - \frac{\lambda_2}{p}\right) / \left(\frac{\lambda_2}{p}\right)^2 = \frac{p(p-\lambda_2)}{\lambda_2^2}$$

In the second model,

$$\begin{aligned} \text{Var}(Y) &= E(\text{Var}(Y|N)) + \text{Var}(E(Y|N)) \\ &= E(Np(1-p)) + \text{Var}(Np) \\ &= \frac{p(1-p)}{\lambda_2} + \frac{p^2(1-\lambda_2)}{\lambda_2^2} \\ &= \frac{p\lambda_2 - p^2\lambda_2 + p^2(1-\lambda_2)}{\lambda_2^2} \\ &= \frac{p\lambda_2 - 2p^2\lambda_2 + p^2}{\lambda_2^2} \end{aligned}$$

Since this is clearly different from the variance of the first model, the marginal distribution of Y must be different. This can also be seen from the fact that the "variance" of Model 1 obtained from plugging in the parameters of 1 could be negative.

3. Find the MLE and a MOM estimator for μ in Model 1.

Since the First Success is a Natural Exponential Family (see section 5), $\hat{\mu}_{\text{MLE}} = \bar{Y}$. Since $E(Y) = \mu$, the Method of Moments gives us $\hat{\mu}_{\text{MOM}} = \bar{Y}$.

4. Find the MLE and a MOM estimator for μ in Model 2. How do they compare to the estimators from 3?

Using the PMF of a binomial distribution, $P(Y_i = y_i|N_i) = \binom{N_i}{y_i}p^{y_i}(1-p)^{N_i - y_i}$. Therefore, the joint distribution is:

$$P(Y_i = y_i, N_i = n_i) = \binom{n_i}{y_i}p^{y_i}(1-p)^{n_i-y_i}\lambda_2(1-\lambda_2)^{n_i-1}$$

Then, the likelihood is

$$\begin{aligned}
L(p, \lambda_2; \vec{y}, \vec{n}) &= \prod_{i=1}^n p^{y_i}(1-p)^{n_i-y_i}\lambda_2(1-\lambda_2)^{n_i-1}\\
&= \left(\prod_{i=1}^n p^{y_i}(1-p)^{n_i-y_i}\right)\left(\prod_{i=1}^n \lambda_2(1-\lambda_2)^{n_i-1}\right)\\
\implies \ell(p, \lambda_2; \vec{y}, \vec{n}) &= \left(\sum_{i=1}^n y_i\log(p) + (n_i - y_i)\log(1-p)\right) + \left(\sum_{i=1}^n \log(\lambda_2) + (n_i - 1)\log(1-\lambda_2)\right)
\end{aligned}$$

Since these are separable, we can maximize each part:

$$\begin{aligned}
\frac{\partial \ell(p, \lambda_2; \vec{y}, \vec{n})}{\partial p} &= \frac{\sum_{i=1}^n y_i}{\hat{p}} - \frac{\sum_{i=1}^n n_i - y_i}{1-\hat{p}} = 0 \implies \hat{p} = \frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n n_i}\\
\frac{\partial \ell(p, \lambda_2; \vec{y}, \vec{n})}{\partial \lambda_2} &= \frac{n}{\hat{\lambda}_2} - \frac{\sum_{i=1}^n (n_i - 1)}{1-\hat{\lambda}_2} = 0 \implies \hat{\lambda}_2 = \frac{n}{\sum_{i=1}^n n_i}
\end{aligned}$$

Then, by invariance of the MLE and the results from 1, $\hat{\mu} = \frac{\hat{p}}{\hat{\lambda}_2} = \frac{\sum_{i=1}^n y_i}{n}$. The method of moments estimator is still $\bar{Y}$. Therefore, using either model and either the MLE or MOM estimator, our estimator for μ is always $\bar{Y}$.

In the Ebola data, our estimates are $\hat{\mu} = 357$, $\bar{n} = 809$, $\hat{p} = 0.44$.

```r
ebola <- read.csv("data/ebola.csv")
round(c("Average deaths" = mean(ebola$Deaths),
        "Average cases" = mean(ebola$Cases),
        "Fatality rate" = sum(ebola$Deaths) / sum(ebola$Cases)), 2)
```

```
## Average deaths  Average cases  Fatality rate
##         356.93         808.58           0.44
```

Section 3: Asymptotics

Asymptotic tools

Law of Large Numbers (LLN): For i.i.d. $Y_i \sim [\mu, \sigma^2]$, $\bar{Y}_n = \frac{1}{n}\sum_{i=1}^n Y_i \xrightarrow{p} \mu$ as $n \to \infty$.

- Note: The moment must exist.
- Strategy: This shows the sample mean of any power of the data converges to its sample moment.
- Example: $\frac{1}{n}\sum_{i=1}^n Y_i^2 \xrightarrow{p} E(Y_i^2)$.

Central Limit Theorem (CLT): For i.i.d. $Y_i \sim [\mu, \sigma^2]$,

$$\sqrt{n}\left(\frac{\bar{Y}_n - \mu}{\sigma}\right) \xrightarrow{d} \mathcal{N}(0, 1)$$

- Strategy: This is often used in combination with Slutsky's Theorem for consistent estimators of σ.
- Strategy: Asymptotic distributions of sample moments can be obtained by plugging in $E(Y^k)$ and $\text{Var}(Y^k)$. These can then be transformed as necessary with the Delta Method.

Slutsky's Theorem: If $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{d} c$ for a constant c, (1) $X_n + Y_n \xrightarrow{d} X + c$, (2) $X_n Y_n \xrightarrow{d} cX$, (3) $X_n / Y_n \xrightarrow{d} X/c$ if $c \neq 0$.

- Note: $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{d} Y$ does not imply $X_n + Y_n \xrightarrow{d} X + Y$ in general (but it does if X_n and Y_n are independent).
- Strategy: Useful when part of the expression involves a sum divided by n since LLN gives convergence to a constant.

Continuous Mapping Theorem (CMT): (1) If $X_n \xrightarrow{d} X$, $g(X_n) \xrightarrow{d} g(X)$. (2) If $X_n \xrightarrow{p} X$, $g(X_n) \xrightarrow{p} g(X)$.

- Strategy: Useful for evaluating convergence of non-standard quantities of distributions (see Invariance of the MLE).

Delta Method: If $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, \omega^2)$ as $n \to \infty$ and g is a differentiable function,

$$\sqrt{n}(g(\hat{\theta}) - g(\theta)) \xrightarrow{d} \mathcal{N}(0, |g'(\theta)|^2 \omega^2)$$

- Strategy: Useful for creating asymptotic confidence intervals for non-standard quantities of distributions.
- Result: Suppose X has the continuous CDF $F(x)$. Then, the p^{th} sample quantile $\hat{Q}(p)$ has the asymptotic distribution $\sqrt{n}(\hat{Q}(p) - Q(p)) \xrightarrow{d} \mathcal{N}\left(0, \frac{p(1-p)}{(f(Q(p)))^2}\right)$.

Score and Fisher

Score: $s(\theta; y) = \ell'(\theta; y)$.

- Fact: $E(s(\theta^*; Y)) = 0$ where θ^* is the true θ for the model and the expectation is with respect to the true model.
- **Information equality**: $-E(s'(\theta^*; Y)) = \text{Var}(s(\theta^*; Y))$ under regularity conditions.
- Fact: $E(s(\hat{\theta}; y)) = 0$ under regularity conditions.

Regularity conditions:

1. The data is i.i.d. $f_\theta(y)$.
2. The support does not depend on θ.
3. $\frac{\partial^3}{\partial \theta^3} f_\theta(y)$ exists.
4. θ^* is not on the boundary of the parameter space.
5. We can differentiate under the integral sign.

Fisher information: $\mathcal{I}_{\vec{Y}}(\theta^*) = \text{Var}(s(\theta^*; \vec{Y}))$.

- Fact: For i.i.d. data, $\mathcal{I}_{\vec{Y}}(\theta^*) = n\mathcal{I}_{Y_1}(\theta^*)$.
- Fact: If regularity conditions hold, $\mathcal{I}_{\vec{Y}}(\theta^*) = -E(s'(\theta^*; \vec{Y}))$.
- **Fisher transformation**: If g is one-to-one and differentiable, $\mathcal{I}(g(\theta)) = \frac{\mathcal{I}(\theta)}{(g'(\theta))^2}$.
- Strategy: Use Fisher information to find asymptotic distributions of MLEs.

Cramer-Rao lower bound: Under regularity conditions, if $\hat{\theta}$ is unbiased for θ, $\text{Var}(\hat{\theta}) \geq \frac{1}{n\mathcal{I}_{Y_1}(\theta^*)}$.

- Equivalently: This is a bound on $\text{MSE}(\hat{\theta})$ since $\hat{\theta}$ is unbiased.

- **Cramer-Rao lower bound for (possibly) biased estimators**: $\text{Var}(\hat{\theta}) \geq \frac{|g'(\theta^*)|^2}{n\mathcal{I}_{Y_1}(\theta^*)}$ where $g(\theta^*) = E(\hat{\theta})$.

MLE consistency: With regularity conditions and a correctly specified model, the MLE is consistent: $\hat{\theta} \xrightarrow{p} \theta^*$.

MLE asymptotics: With regularity conditions and a correctly specified model, the MLE has the asymptotic distribution:

$$\sqrt{n}(\hat{\theta} - \theta^*) \xrightarrow{d} \mathcal{N}\left(0, \frac{1}{\mathcal{I}_{Y_1}(\theta^*)}\right)$$

- Fact: The MLE achieves the CRLB.
- Example: Let $Y_1, ..., Y_n \sim \text{Pois}(\lambda)$. Then, with $\hat{\lambda} = \bar{Y}$, by the CLT,

  $$\sqrt{n}(\hat{\lambda} - \lambda) \xrightarrow{d} \mathcal{N}(0, \lambda)$$

  Applying the transformation $g(\lambda) = \sqrt{\lambda}$, the Delta Method gives

  $$\sqrt{n}(\sqrt{\hat{\lambda}} - \sqrt{\lambda}) \xrightarrow{d} \mathcal{N}\left(0, \frac{\lambda}{4(\sqrt{\lambda})^2}\right) \iff \sqrt{\sum_{i=1}^n Y_i} - \sqrt{n\lambda} \xrightarrow{d} \mathcal{N}(0, 1/4)$$

  As an approximation, this says

  $$\text{Var}\left(\sqrt{\sum_{i=1}^n Y_i} - \sqrt{n\lambda}\right) = \text{Var}\left(\sqrt{\sum_{i=1}^n Y_i}\right) = 1/4$$

  Since the sum of Poisson random variables is Poisson, this says that for sufficiently large λ, with $W \sim \text{Pois}(\lambda)$, $\text{Var}(\sqrt{W}) = 1/4$.

Problem 2

In this problem, we will find the variance stabilizing transformation of the Binomial distribution. Let $Y \sim \text{Bin}(n, p)$.

1. Find the asymptotic distribution of the MLE $\hat{p}$ as $n \to \infty$.

Treating Y as a sum of i.i.d. $\text{Bern}(p)$ random variables, the Central Limit Theorem gives:

$$\sqrt{n}(\hat{p} - p) \to \mathcal{N}(0, p(1-p))$$

2. Find the Fisher information of p for a single $\text{Bern}(p)$. What is the Fisher information for a one-to-one differentiable transformation $\theta = g(p)$?

If $Y_1 \sim \text{Bern}(p)$, the score is $s(p; y_1) = \frac{y_1}{p} - \frac{1-y_1}{1-p} = \frac{y_1 - p}{p(1-p)}$, so

$$\mathcal{I}_{Y_1}(p) = \text{Var}(s(p; Y_1)) = \frac{\text{Var}(Y_1)}{p^2(1-p)^2} = \frac{1}{p(1-p)}$$

For the transformation,

$$\mathcal{I}_{Y_1}(\theta) = \frac{\mathcal{I}_{Y_1}(p)}{g'(p)^2} = \frac{1}{g'(p)^2 p(1-p)}$$

3. Show that if $\theta = g(p) = \arcsin(\sqrt{p})$, $\mathcal{I}_1(\theta)$ is a constant.

$$g'(p) = \frac{1}{2\sqrt{p}}\frac{1}{\sqrt{1-p}} \implies \mathcal{I}_{Y_1}(\theta) = \frac{1}{g'(p)^2 p(1-p)} = 4$$

4. Write an approximate distribution for a function g of p and n such that for large n the variance of $g(\hat{p})$ is constant.

Let $g(p) = \arcsin(\sqrt{p})$. By the Delta Method, we have

$$\sqrt{n}(g(\hat{p}) - g(p)) \to \mathcal{N}\left(0, 1/4\right)$$

As an approximation, this says that for large n

$$\arcsin(\sqrt{\hat{p}}) \dot{\sim} \mathcal{N}\left(\arcsin(\sqrt{p}), \frac{1}{4n}\right) \iff \sqrt{n}\arcsin(\sqrt{Y/n}) \dot{\sim} \mathcal{N}\left(\arcsin(\sqrt{p}), 1/4\right)$$

The table below shows $\text{Var}(\hat{p})$ and $\text{Var}(\sqrt{n}\arcsin(\sqrt{Y/n}))$ from 10^5 simulations with $n = 10^4$ and various values of p.

	0.01	0.05	0.25	0.5	0.75	0.95	0.99
Unstabilized	0.0099	0.0474	0.1880	0.2521	0.1869	0.0477	0.0100
Variance Stabilized	0.2520	0.2508	0.2499	0.2504	0.2496	0.2496	0.2502

Section 4: Intervals

Interval estimator: An interval estimator is an interval $C(\vec{Y}) = [L(\vec{Y}), U(\vec{Y})]$ with $L(\vec{Y}) \leq U(\vec{Y})$.

Coverage: If an interval $C(\vec{Y})$ for θ contains θ ($\theta \in C(\vec{Y})$), the interval covers θ.

Coverage probability: $P(\theta \in C(\vec{Y})|\theta)$.

- Note: The coverage probability is a function of θ.

Confidence interval: An interval estimator with coverage probability at least $1-\alpha$ for all possible values of θ is a $100(1-\alpha)\%$ confidence interval.

- Interpretation: The probability that the random interval generated by repeated draws of the data contains the fixed θ is $1-\alpha$.
- Biohazard: Not the probability a particular interval contains θ.
- Note: The confidence interval is not unique; there could be multiple intervals with coverage probability $1-\alpha$.

Pivotal interval: A pivot is a function of the data and the parameter(s) whose distribution does not depend on the parameter(s). These can be used to construct intervals.

- Strategy: The Normal, t, Gamma, and χ^2 distributions are commonly used as pivots.
- Example: For $Y_1, ..., Y_n \sim \mathcal{N}(\mu, \sigma^2)$ with both μ and σ^2 unknown, the quantity $\frac{\bar{Y}-\mu}{\sqrt{\hat{\sigma}^2/n}}$ has a t_{n-1} distribution, so we can construct an exact 95% confidence interval for μ as:

$$P\left(-Q_{t_{n-1}}(0.975) \leq \frac{\bar{Y}-\mu}{\sqrt{\hat{\sigma}^2/n}} \leq Q_{t_{n-1}}(0.975)\right) = 0.95$$

$$\implies P\left(\bar{Y} - Q_{t_{n-1}}(0.975)\sqrt{\hat{\sigma}^2/n} \leq \mu \leq \bar{Y} + Q_{t_{n-1}}(0.975)\sqrt{\hat{\sigma}^2/n}\right) = 0.95$$

so the interval is $\left[\bar{Y} - Q_{t_{n-1}}(0.975)\sqrt{\hat{\sigma}^2/n}, \bar{Y} + Q_{t_{n-1}}(0.975)\sqrt{\hat{\sigma}^2/n}\right]$.

Asymptotic interval: Using the asymptotic distribution of $\hat{\theta}$, we can find an approximate pivot.

- Example: Let $Y_1, ..., Y_n \sim \text{Expo}(\lambda)$. Then, by the Central Limit Theorem

  $$\frac{\sqrt{n}(\bar{Y} - 1/\lambda)}{\sqrt{1/\lambda^2}} \xrightarrow{d} \mathcal{N}(0, 1)$$

  Then, we can construct an approximate interval:

  $$\begin{aligned} 0.95 &= P(-Q_{\mathcal{N}(0,1)}(0.975) \le \sqrt{n}(\bar{Y}\lambda - 1) \le Q_{\mathcal{N}(0,1)}(0.975)) \\ &= P(-Q_{\mathcal{N}(0,1)}(0.975)/\sqrt{n} + 1 \le \bar{Y}\lambda \le Q_{\mathcal{N}(0,1)}(0.975)/\sqrt{n} + 1) \\ &= P\left(\frac{-Q_{\mathcal{N}(0,1)}(0.975)/\sqrt{n} + 1}{\bar{Y}} \le \lambda \le \frac{Q_{\mathcal{N}(0,1)}(0.975)/\sqrt{n} + 1}{\bar{Y}}\right) \end{aligned}$$

  so the interval is

  $$\left[\frac{-Q_{\mathcal{N}(0,1)}(0.975)/\sqrt{n} + 1}{\bar{Y}}, \frac{Q_{\mathcal{N}(0,1)}(0.975)/\sqrt{n} + 1}{\bar{Y}}\right]$$

Section 5: Sufficient statistics and EFs

Sufficient statistics: For $Y_1, ..., Y_n$ from $F_{\vec{Y}|\theta}$, a statistic $T(\vec{Y})$ is sufficient for θ if the conditional distribution $Y_1, ..., Y_n | T(\vec{Y})$ does not depend on θ.

- Note: Sufficient statistics are not unique.

Factorization criterion: $T(\vec{Y})$ is sufficient for θ iff we can factor the joint PDF/PMF as $f_{\vec{Y}}(\vec{y}|\theta) = g(T(\vec{y}), \theta)h(\vec{y})$.

- Strategy: Since $h(y)$ does not depend on θ, $L(\theta; \vec{y}) \propto g(T(\vec{y}), \theta)$, so the likelihood can be written purely in terms of $T(\vec{y})$ and θ.

Rao-Blackwell: If T is a sufficient statistic and $\hat{\theta}$ is an estimator for θ, $\text{MSE}(\hat{\theta}_{RB}) = \text{MSE}(E(\hat{\theta}|T)) \le \text{MSE}(\hat{\theta})$.

- Note: Rao-Blackwell only improves variance, not bias.
- Rao-Blackwell will not change $\hat{\theta}$ that are already functions of T. In particular, it will not improve the MLE.

Natural Exponential Family (NEF): Y follows an NEF if its PDF is of the form

$$\exp\Big(y\theta - \Psi(\theta)\Big)h(y)$$

- Examples: Poisson, Binomial (n fixed), Negative Binomial (r fixed), Normal (σ^2 known), Gamma (a known).

NEF Properties:

1. $E(Y) = \Psi'(\theta)$, $\text{Var}(Y) = \Psi''(\theta)$, the MGF is $E(e^{tY}) = \exp\Big(\Psi(\theta + t) - \Psi(\theta)\Big)$.
2. $\bar{Y}$ is a sufficient statistic for θ.
3. The MLE for $E(Y)$ is $\bar{Y}$.
4. The Fisher information of a single observation is $\mathcal{I}_{Y_1}(\theta) = \Psi''(\theta)$.

- Example: For $Y \sim \text{FS}(\lambda)$,

$$\begin{aligned}
P(Y=y) &= \lambda(1-\lambda)^{y-1} \\
&= \exp\Big(\log(\lambda) + (y-1)\log(1-\lambda)\Big) \\
&= \exp\Big(y\log(1-\lambda) - \log(1-\lambda) + \log(\lambda)\Big) \\
&= \exp\left(y\log(1-\lambda) - \log\left(\frac{1-\lambda}{\lambda}\right)\right) \\
&= \exp\Big(y\theta - \Psi(\theta)\Big)h(y)
\end{aligned}$$

  where $h(y) = 1$, $\theta = \log(1-\lambda)$, and $\Psi(\theta) = \log\left(\frac{\exp(\theta)}{1-\exp(\theta)}\right)$.

Exponential family (EF): Y follows an EF if its PDF is of the form

$$\exp\Big(T(y)\theta - \Psi(\theta)\Big)h(y)$$

for some function T.

- Examples: Normal with μ and σ^2 unknown, Weibull.

Section 6: Regression

Predictive regression: Attempting to predict an outcome variable Y from a vector of covariates $\overrightarrow{X}$: $\mu(\vec{x}) = \mu(Y|\overrightarrow{X} = \vec{x})$.

- In words: Expected value of the outcome given the predictors.

Regression error: $U(\vec{x}) = Y - \mu(\vec{x})$.

- In words: The difference between the observed outcome and the predicted outcome.
- Note: This is still a random variable where the randomness comes from Y.
- Note: This is unobservable because this would require knowing the true $\mu(\vec{x})$.
- Fact: $E(U(\overrightarrow{X})) = 0$.
- Fact: $\text{Cov}(U(\overrightarrow{X}), \overrightarrow{X}) = 0$.

Homoscedasticity: $\text{Var}(U_j|\overrightarrow{X} = \vec{x}) = \sigma^2$ for all j.

- Alternative: **Heteroscedastic** (different variances for different errors).

Linear regression: The regression function is linear in the parameters: $\mu(\vec{x}) = \theta_0 + \theta_1 x_1 + ... + \theta_k x_k$.

- Note: Linear in the parameters, not the predictors.

Residual: $\hat{U_j} = Y_j - \hat{\theta}\vec{x_j}$.

- Note: We *can* compute this from the data (contra regression error).

MLE for predictive regression: $\vec{\hat{\theta}}_{\text{MLE}} = \text{argmax}_{\vec{\theta}} \prod_{i=1}^n f(y_i|\overrightarrow{X_i} = \vec{x_i}, \vec{\theta})$.

- Example: For $Y_i|X_j = \theta X_j + \epsilon_j,\ \epsilon_j \sim \mathcal{N}(0, \sigma^2)$,

$$\hat{\theta}_{\text{MLE}} = \frac{\sum_{i=1}^n Y_i x_i}{\sum_{i=1}^n x_i^2}$$

Least squares estimator: $\hat{\theta}_{\text{LS}} = \text{argmin}_\theta \sum_{i=1}^n (y_i - \vec{x_i} \cdot \vec{\theta})^2$.

- Example: $\hat{\theta}_{\text{LS}} = \hat{\theta}_{\text{MLE}}$ for homoscedastic Normal errors.

Logistic regression:

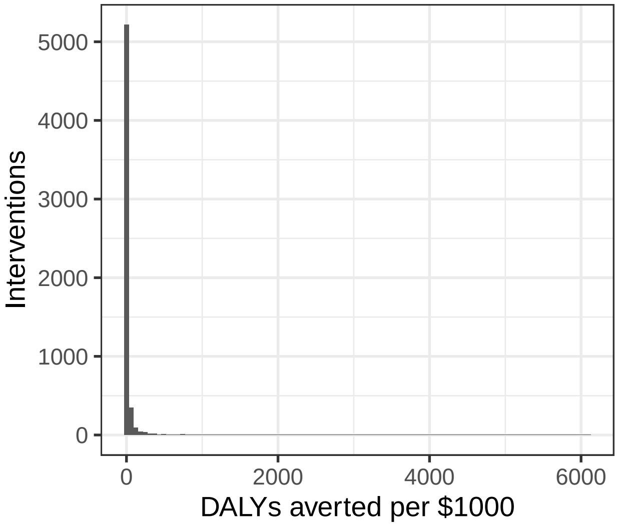
$$P(Y=1|\vec{X}=\vec{x},\vec{\theta}) = \frac{\exp(\theta_0 + \sum_{k=1}^{K} x_k\theta_k)}{1+\exp(\theta_0 + \sum_{k=1}^{K} x_k\theta_k)}$$

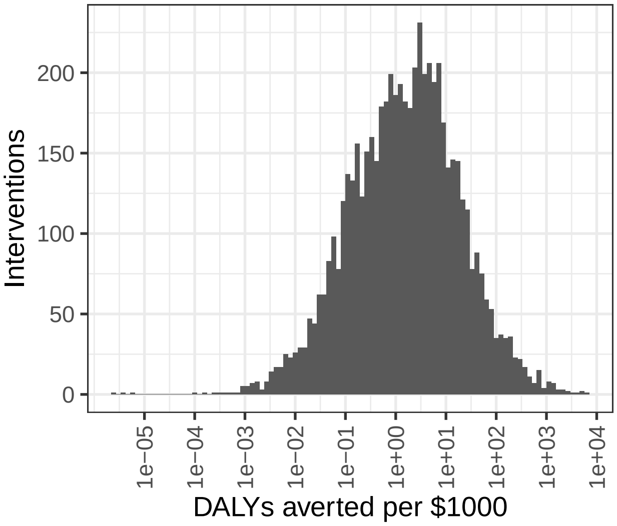
Descriptive regression: If X, Y have a joint distribution, the descriptive regression is $\beta_{Y\sim X} = \frac{\text{Cov}(X,Y)}{\text{Var}(X)}$.

- Note: Summary of the relationship rather than conditioning on X to predict Y.
- Fact: $(E(Y) - \beta_{Y\sim X}E(X), \beta_{Y\sim X})$ are the (a,b) that minimize $E((Y-(a+bX))^2)$.

Problem 3

Significant work in economics and global health has gone into determining which healthcare interventions prevent the most suffering per dollar spent. The metric of choice for these evaluations is the DALY, a disability adjusted life year. Because some global health interventions save lives while others improve lives, the DALY attempts to provide a standardized unit of comparison. For example, antenatal syphilis screening in Uganda in 2013 was estimated to avert 124 DALYs per \$1000 USD. The Tufts Medical Center Cost-Effectiveness Analysis Registry aggregates academic literature on healthcare interventions and standardizes the results to compare different interventions. This literature is notoriously variable and unstandardized, so the specific metrics for each intervention might not be accurate, but the general trends provide some insight. Here, we will try to compare the median cost effectiveness of an intervention in a low-income country (GNI per capita below \$1085) versus a high-income country (GNI per capita above \$13205).





County	Adjusted DALYs averted per \$1000	Studies
Zimbabwe	46.09	101
Guinea	17.68	23
Benin	14.04	22
Papua New Guinea	11.71	11
Madagascar	11.56	27

County	Adjusted DALYs averted per \$1000	Studies
Israel	0.01	15

County	Adjusted DALYs averted per $1000	Studies
Qatar	0.022	2
Denmark	0.03	17
Australia	0.032	236
Netherlands	0.034	30

1. Suppose we have two independent groups each with many independent observations: $X_{1,1}, ..., X_{1,n_1} \sim \mathcal{LN}(\mu_1, \sigma_1^2)$ and $X_{2,1}, ..., X_{2,n_2} \sim \mathcal{LN}(\mu_2, \sigma_2^2)$ where n_1 and n_2 are large. Let $Y_{i,j} = \log(X_{i,j})$, so the $Y_{i,j}$ are distributed Normally. Write the distribution of $\bar{Y}_1 - \bar{Y}_2$, and use this to write an approximate 95% confidence interval for $\mu_1 - \mu_2$. Assume all the parameters are unknown.

$\bar{Y}_1 \sim \mathcal{N}(\mu_1, \sigma_1^2/n_1)$ and $\bar{Y}_2 \sim \mathcal{N}(\mu_2, \sigma_2^2/n_2)$, so $\bar{Y}_1 - \bar{Y}_2 \sim \mathcal{N}(\mu_1 - \mu_2, \sigma_1^2/n_1 + \sigma_2^2/n_2)$. This is equivalent to

$$\frac{(\bar{Y}_1 - \bar{Y}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \sim \mathcal{N}(0, 1)$$

and since the sample variance of the Y_1s $S_1^2 \to \sigma_1^2$, by Slutsky's Theorem

$$\frac{(\bar{Y}_1 - \bar{Y}_2) - (\mu_1 - \mu_2)}{\sqrt{S_1^2/n_1 + S_2^2/n_2}} \xrightarrow{d} \mathcal{N}(0, 1)$$

for large n. Let $S' = S_1^2/n_1 + S_2^2/n_2$. Thus, the approximate interval is

$$\left[\bar{Y}_1 - \bar{Y}_2 - Q(0.975)\sqrt{S'}, \bar{Y}_1 - \bar{Y}_2 + Q(0.975)\sqrt{S'}\right]$$

2. Let $M_{X,1}$ be the median of the $X_{1,i}$s and $M_{X,2}$ be the median of the $X_{2,i}$s. Construct an approximate 95% confidence interval for $M_{X,1}/M_{X,2}$.

Let $M_{Y,1}$ be the median of the $Y_{1,i}$s and $M_{Y,2}$ be the median of the $Y_{2,i}$s. Since the median equals the mean in a Normal distribution, our interval from (1) is also an interval for the difference in medians of the Y_is. Exponentiating both sides gives:

$$P\left[e^{\left(\bar{Y}_1 - \bar{Y}_2 - Q(0.975)\sqrt{S'}\right)} \leq e^{(M_{Y,1} - M_{Y,2})} \leq e^{\left(\bar{Y}_1 - \bar{Y}_2 + Q(0.975)\sqrt{S'}\right)}\right] \approx 0.95$$

$$\implies P\left[e^{\left(\bar{Y}_1 - \bar{Y}_2 - Q(0.975)\sqrt{S'}\right)} \leq \frac{M_{X,1}}{M_{X,2}} \leq e^{\left(\bar{Y}_1 - \bar{Y}_2 + Q(0.975)\sqrt{S'}\right)}\right] \approx 0.95$$

Therefore, the interval is

$$\left[e^{\left(\bar{Y}_1 - \bar{Y}_2 - Q(0.975)\sqrt{S'}\right)}, e^{\left(\bar{Y}_1 - \bar{Y}_2 + Q(0.975)\sqrt{S'}\right)}\right]$$

```
##      P-value Lower bound Upper bound
##      0.00000    53.40359    79.60798
```

For the actual data, this 95% confidence interval is 53.4 to 79.6, so $1000 spent on the median intervention in a low-income country averts 53 to 80 times as many DALYs as that $1000 spent on the median intervention in a high-income country. Put another way, the cost to save an entire life in a low-income country is about the same as the cost to save one year of healthy life in a high-income country.

Break

Section 7: Hypothesis testing

Null and alternative hypotheses: Consider a partition of the parameter space Θ into two disjoint sets Θ_0 and Θ_1 such that $\Theta = \Theta_0 \cup \Theta_1$. Then, $H_0 : \theta \in \Theta_0$ and $H_a : \theta \in \Theta_1$.

- Note: The null hypothesis is what we want to disprove.
- **One-sided**: $H_0 : \theta \leq \theta_0$ vs. $H_a : \theta > \theta_0$.
- **Two-sided**: $H_0 : \theta = \theta_0$ vs. $H_a : \theta \neq \theta_0$.
- **Simple hypothesis**: $\Theta_0 = \{\theta_0\}$.
- **Composite hypothesis**: Θ_0 is an interval or intervals.
- Example: Let $Y_1, ..., Y_n \sim \mathcal{N}(\mu, \sigma^2)$ and consider the hypotheses $H_0 : \mu = 0$ vs. $H_a : \mu \neq 0$ (this is a two-sided simple hypothesis).

Rejection region: A subset R of the range of data $\vec{y}$ such that we reject H_0 if $\vec{y} \in R$ and fail to reject H_0 if $y \notin R$.

- Strategy: Find a **test statistic** $t(\vec{Y})$ and use the rejection region $R = \{\vec{y} : t(\vec{y}) > c\}$ (one-sided) or $R = \{\vec{y} : t(\vec{y}) < c_L \text{ or } t(\vec{y}) > c_U\}$ for **critical values** c, c_L, c_U.
- Strategy: The test statistic should have a known distribution (commonly Normal, student-t, Gamma, or Binomial).
- Note: Iff the data is in the rejection region, the corresponding confidence interval will not contain θ_0.
- Example: Using the Y_i above, consider the test statistic $t(\vec{Y}) = \frac{\bar{Y}}{\sqrt{\hat{\sigma}^2/n}}$. Under the null, this has a t_{n-1} distribution, so the critical value is $c_U = -c_l = Q_{t_{n-1}}(0.975)$, and the rejection region is $R = \{\vec{y} : \left|\frac{\bar{y}}{\sqrt{\hat{\sigma}^2/n}}\right| > Q_{t_{n-1}}(0.975)\}$.

Type I error: Rejecting the null when the null is true.

- Strategy: Controlled by α level of the test.

Type II error: Retaining the null when the null is false.

Power: Probability of rejecting the null, $\beta(\theta) = P(\vec{Y} \in R|\theta)$.

- Strategy: If $\theta \in \Theta_1$, $\beta(\theta) = 1 - P(\text{Type II error})$.
- Strategy: If $\theta \in \Theta_0$, $\beta(\theta) = P(\text{Type I error})$, which we try to fix at α.

P-value: If R_α is the rejection region for a test with a Type I error rate of α, the p-value is the smallest α at which we reject H_0: $p = \inf\{\alpha : t(\vec{y}) \in R_\alpha\}$.

- In words: The probability of obtaining data as or more extreme than the observed data under the null.
- Strategy: For a particular α, if the p-value is less than α, we reject H_0.
- Biohazard: Not the probability H_0 is true.
- Fact: In a two-sided test of $H_0 : \theta = \theta_0$ vs. $H_a : \theta \neq \theta_0$ with a continuously distributed $t(\vec{Y})$, the p-value is uniformly distributed under the null.
- p-hacking: Testing many hypothesis increases the probability of observing a low p-value by chance.
- Example: With the Y_i above, the p-value would be $2F_{t_{n-1}}\left(\frac{-|\bar{y}|}{\sqrt{\hat{\sigma}^2/n}}\right)$.

Asymptotic tests: For sufficiently large n and $H_0 : \theta = \theta_0$, $H_a : \theta \neq \theta_0$:

1. **Wald test**: $\sqrt{n\mathcal{I}_1(\theta_0)}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0,1)$, so reject the null if

$$\left|\sqrt{n\mathcal{I}_1(\theta_0)}(\hat{\theta} - \theta_0)\right| > Q_{\mathcal{N}(0,1)}(1-\alpha/2)$$

2. **Score test**: $\frac{S(\vec{Y},\theta_0)}{\sqrt{n\mathcal{I}_1(\theta_0)}} \xrightarrow{d} \mathcal{N}(0,1)$, so reject the null if

$$\left|\frac{S(\vec{Y},\theta_0)}{\sqrt{n\mathcal{I}_1(\theta_0)}}\right| > Q_{\mathcal{N}(0,1)}(1-\alpha/2)$$

3. **Likelihood ratio test**: Let

$$\Lambda(\vec{Y}) = 2\log\left(\frac{L(\hat{\theta};\vec{Y})}{L(\theta_0;\vec{Y})}\right)$$

where $\hat{\theta}$ is the MLE for θ. Under regularity conditions, $\Lambda(\vec{Y}) \xrightarrow{d} \chi^2_1$, so reject the null if

$$2\log\left(\frac{L(\hat{\theta};\vec{Y})}{L(\theta_0;\vec{Y})}\right) > Q_{\chi^2_1}(1-\alpha)$$

- In words: If the likelihood is much higher under $\hat{\theta}$ than under the null, we have evidence to reject the null.

Problem 4

Many development economists have argued that cost-sharing, charging a much-reduced but non-zero price for healthcare resources, is necessary to avoid wasting the resources on people who do not need them. In their 2010 paper "Free Distribution or Cost-Sharing? Evidence from a Randomized Malaria Prevention Experiment," Jessica Cohen and Pascaline Dupas claim to show there is no evidence that cost-sharing reduces wastage of Insecticide Treated Nets (used to prevent malaria). However, they show that cost-sharing does significantly decrease demand for ITNs. Part of the study involved randomizing the cost of ITNs at rural Kenyan health clinics for pregnant women and tracking ITN sales. Four prices were used ($0, $0.15, $0.30, and $0.60) at multiple clinics each. We want to perform a regression for predicting net sales from the price charged to see if there is an association.

1. Let $Y_i|X_i = x_i \sim \mathcal{N}(\theta x_i, \sigma^2_\epsilon)$ for $i \in \{1,...,n\}$. Suppose the data comes in k groups $S_1,...,S_k$ so that all the predictors (the X_i) in a group are the same (define this as $x_i = x'_j$ for $i \in S_j$). Let $\bar{Y}_j$ be the mean of Y_is in group j ($\bar{Y}_j = \frac{1}{n_j}\sum_{i:i\in S_j} Y_i$), $\hat{\sigma}^2_j$ be the MLE variance of the Y_i in group j ($\hat{\sigma}^2_j = \frac{1}{n_j}\sum_{i:i\in S_j}(Y_i - \bar{Y}_j)^2 = \frac{1}{n_j}\sum_{i:i\in S_j}(Y_i^2 - \bar{Y}_j^2)$), and n_j be the size of the group. Show that $(\bar{Y}_1,...,\bar{Y}_k,\hat{\sigma}^2_1,...,\hat{\sigma}^2_k)$ is a sufficient statistic. (Treat the x_i and n_j as known. Also, it doesn't actually matter whether σ^2_ϵ is known or unknown, but you can treat it as known.)

$$
\begin{aligned}
L(\theta; \vec{x}, \vec{y}) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_\epsilon^2}} \exp\left(-\frac{1}{2}\frac{(y_i - \theta x_i)^2}{\sigma_\epsilon^2}\right) \\
&\propto \exp\left(-\frac{1}{2\sigma_\epsilon^2}\sum_{i=1}^{n}(y_i - \theta x_i)^2\right) \\
&= \exp\left(-\frac{1}{2\sigma_\epsilon^2}\sum_{i=1}^{n} y_i^2 - 2\theta x_i y_i + (\theta x_i)^2\right) \\
&= \exp\left(-\frac{1}{2\sigma_\epsilon^2}\sum_{j=1}^{k}\left(\sum_{i:i\in S_j} y_i^2 - 2\theta x_i y_i + (\theta x_i)^2\right)\right) \\
&= \exp\left(-\frac{1}{2\sigma_\epsilon^2}\sum_{j=1}^{k}\left(-2n_j\theta x_j'\bar{y}_i + n_j(\theta x_j')^2 + \sum_{i:i\in S_j} y_i^2\right)\right) \\
&= \exp\left(-\frac{1}{2\sigma_\epsilon^2}\sum_{j=1}^{k}\left(-2n_j\theta x_j'\bar{y}_i + n_j(\theta x_j')^2 + n_j(\bar{y}_j^2 + \hat{\sigma}_j^2)\right)\right)
\end{aligned}
$$

Since the likelihood is written only in terms of known parameters and $(\bar{Y}_1, ..., \bar{Y}_k, \sigma_1^2, ..., \sigma_k^2)$, $(\bar{Y}_1, ..., \bar{Y}_k, \sigma_1^2, ..., \sigma_k^2)$ is a sufficient statistic. In words, this says that in a linear regression, we only need the mean and variance of the outcomes for each predictor group, not all the individual data points.

2. Now consider the model $Y_i|X_i = x_i \sim \mathcal{N}(\theta_1 x_i + \theta_0, \sigma_\epsilon^2)$. When $\theta_1 = 0$, the least squares estimate $\hat{\theta_1}$ has the distribution:

$$\frac{\hat{\theta_1}}{\hat{\sigma}\sqrt{\frac{1}{\sum_{i=1}^{n} x_i^2}}} \sim t_{n-2}, \quad \hat{\sigma}^2 = \frac{1}{n-2}\sum_{i-1}^{n}(y_i - (\hat{\theta_1}x_i + \hat{\theta_0}))^2$$

Show how to conduct a hypothesis test for whether price has any effect on the number of ITNs sold.

Our hypotheses are $H_0 : \theta_1 = 0$ vs. $H_a : \theta_1 \neq 0$. To find the p-value, we would compute $\hat{T} = \frac{\hat{\theta_1}}{\hat{\sigma}\sqrt{\frac{1}{\sum_{i=1}^{n} x_i^2}}}$ for the data and find the p-value with $F(-|T|) + 1 - F(|T|)$, where F is the CDF of the t_{n-2} distribution.

In the actual data, we obtain $\hat{\theta_1} = -0.81$ and a p-value of 9.3×10^{-4}, so we reject the null and conclude that fewer ITNs are sold when the prices are higher.

```
##               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 43.0699643  4.2634450 10.102151 2.218836e-16
## cost        -0.8081551  0.2358893 -3.425993 9.329103e-04
```

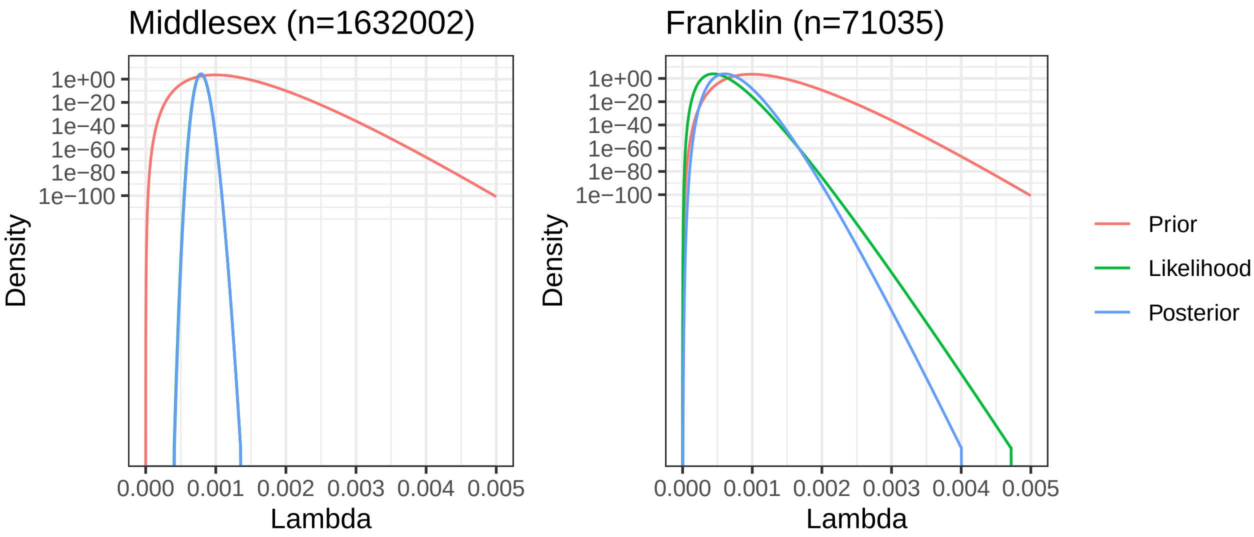
8: Bayesian statistics

Bayes' rule:

$$f(\theta|\vec{y}) = \frac{f(\vec{y}|\theta)f(\theta)}{f(\vec{y}|\theta)} \propto L(\theta; \vec{y})f(\theta)$$

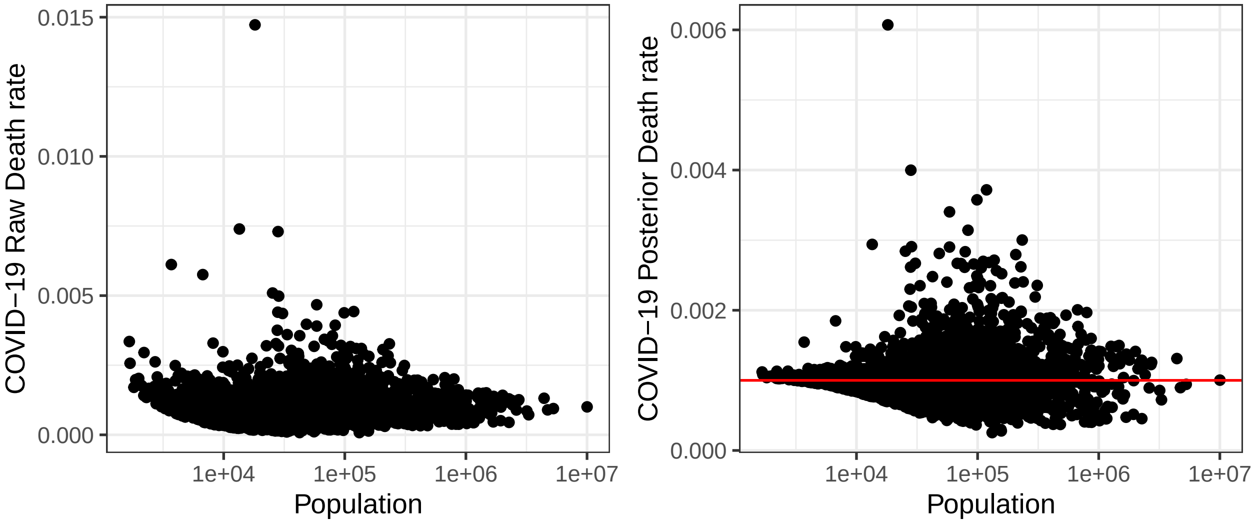
Posterior distribution: The distribution of $\theta|\vec{y}$.

- Note: Compromise between the likelihood and the prior with weighting based on the sample size.
- Example (Nickols section 8): Suppose we model COVID-19 deaths in a particular US county as $Y_i \sim \text{Pois}(c\lambda_i n_i)$ where $c = 3.23$ is the number of years included in the data set, λ_i is the county's annual death rate from COVID-19, and n_i is the county's population. Also, suppose we use the prior $\lambda_i \sim \text{Gamma}(a, b)$ with $b = 10^5$ and $a = 0.001b$.



County	Population	Rate (Unadjusted)
Montour County, Pennsylvania	18145	0.0147
Martinsville city, Virginia	13486	0.0074
Winchester city, Virginia	28122	0.0073
Norton city, Virginia	3696	0.0061
Galax city, Virginia	6725	0.0058

County	Population	Rate (Adjusted)
Montour County, Pennsylvania	18145	0.0061
Winchester city, Virginia	28122	0.004
Potter County, Texas	118527	0.0037
Madison County, Tennessee	98843	0.0036
Newton County, Missouri	58644	0.0034



Posterior predictive distribution: The distribution of $Y_{n+1}|Y_1, ..., Y_n$.

Point estimates:

- **Posterior mean**: $E(\theta|\vec{y})$, minimizes expected square loss $E((\theta-\tilde{\theta})^2|\vec{y})$.
- **Posterior median**: $Q_{\theta|\vec{y}}(0.5)$, minimizes expected absolute loss $E(|\theta-\tilde{\theta}||\vec{y})$.
- **Posterior mode**: $\text{argmax}_\theta f(\theta|\vec{y})$.

Credible interval: $[L(\vec{y}), U(\vec{y})]$ is a $1-\alpha$ credible interval if $P(L(\vec{y}) \leq \theta \leq U(\vec{y})|\vec{y}) = 1-\alpha$.

- Strategy: Find the posterior distribution for θ and use the 2.5th and 97.5th quantiles.

Conjugacies:

- **Beta-Binomial**: Let $\theta \sim \text{Beta}(a,b)$ and $Y_i|\theta \sim \text{Bin}(n_i,\theta)$. Then,
$$\theta|\vec{Y} \sim \text{Beta}(a+\sum_{i=1}^n Y_i, b+\sum_{i=1}^n n_i - \sum_{i=1}^n Y_i)$$
- **Gamma-Poisson-Negative Binomial**: Let $\theta \sim \text{Gamma}(a,b)$ and $Y_i|\theta \sim \text{Pois}(\theta t_i)$. Then, $\theta|\vec{Y} \sim \text{Gamma}(a+\sum_{i=1}^n Y_i, b+\sum_{i=1}^n t_i)$. Also,
$$Y_{n+1}|\vec{Y} \sim \text{NBin}\left(a+\sum_{i=1}^n Y_i, \frac{b+\sum_{i=1}^n t_i}{t_{n+1}+b+\sum_{i=1}^n t_i}\right)$$
- **Normal-Normal**: Let $\mu \sim \mathcal{N}(\mu_0,\sigma_0^2)$ and $Y_i|\mu \sim \mathcal{N}(\mu,\sigma_i^2)$ with $\sigma_0^2,\mu_0,\sigma_i^2$ known. Then, $\mu|\vec{Y} \sim \mathcal{N}(\mu_n,\tau_n^2)$ where
$$\tau_n^2 = \frac{1}{\frac{1}{\sigma_0^2}+\sum_{i=1}^n \frac{1}{\sigma_i^2}},\ \mu_n = \tau_n^2\left(\frac{\mu_0}{\sigma_0^2}+\sum_{i=1}^n \frac{Y_i}{\sigma_i^2}\right)$$
Also,
$$Y_{n+1}|\vec{Y} \sim \mathcal{N}(\mu_n, \tau_n^2+\sigma_{n+1}^2)$$

9: Risk, James Stein

Loss function: A convex function $C(\theta,\hat{\theta}) \geq 0$ with the property that $C(x,x)=0$ for all x.

- In words: How bad is it to guess $\hat{\theta}$ when the true value is θ.

Risk function: $r_{\hat{\theta}}(\theta) = E(C(\theta,\hat{\theta})|\theta)$.

Admissibility: An estimator $\hat{\theta}$ is inadmissible if there is some other $\tilde{\theta}$ such that, for all θ, $r_{\tilde{\theta}}(\theta) \leq r_{\hat{\theta}}(\theta)$ with strict inequality for some θ. Otherwise, the estimator is admissible.

James-Stein Estimator: Let independent $Y_i|\mu_i,V \sim \mathcal{N}(\mu_i,V)$ for $i=1,...,k$ with $k \geq 3$ and V known. Let $\mu=(\mu_1,...,\mu_k)$ and $\hat{\mu}=(Y_1,...,Y_k)$. Using squared loss $C(\mu,\hat{\mu})=\sum_{i=1}^k(\mu_i-\hat{\mu}_i)^2$, $\hat{\mu}$ is inadmissible. $\hat{\mu}$ can be dominated by
$$\hat{\mu}_{i,JS} = \left(1-\frac{(k-2)V}{S}\right)Y_i$$
where $S=\sum_{i=1}^k Y_i^2$.

- Example: Estimates of entire-season batting averages as predicted by early-season batting averages can be improved by shrinking estimates towards the grand mean.

10: Sampling

Survey sampling: Suppose the population is of size N and let the variables of interest $y_1, ...y_N$ be fixed. The population estimands are $\mu = \frac{1}{N}\sum_{i=1}^N y_i$, and $\sigma^2 = \frac{1}{N}\sum_{i=1}^N (y_i - \mu)^2$.

Simple random sample (SRS) with replacement: Sample n items (with replacement) from the N total and call these sampled observations $Y_1, ..., Y_n$. Consider the estimator $\hat{\mu} = \bar{Y}$.

- Fact: $E(Y_i) = \mu$, $\text{Var}(Y_i) = \sigma^2$, $\text{Cov}(Y_i, Y_j) = 0$.
- Fact: $E(\hat{\mu}) = \mu$, $\text{Var}(\hat{\mu}) = \sigma^2/n$.

SRS without replacement: Sample n items (without replacement) from the N total and call these sampled observations $Y_1, ..., Y_n$. Consider the estimator $\hat{\mu} = \bar{Y}$.

- Fact: $E(Y_i) = \mu$, $\text{Var}(Y_i) = \sigma^2$, $\text{Cov}(Y_i, Y_j) = -\sigma^2/(N-1)$.
- Fact: $E(\hat{\mu}) = \mu$, $\text{Var}(\hat{\mu}) = \frac{\sigma^2}{n}\frac{N-n}{N-1}$ ($\frac{N-n}{N-1}$ is the finite population correction).

Stratified sampling: Divide the population into K strata such that $\sum_{i=1}^K N_i = N$ where the N_i are known, and let μ_i and σ_i^2 be the mean and variance of stratum i. Take a sample of n_i from stratum i for each i. Consider the estimator of the overall mean $\hat{\mu} = \sum_{i=1}^K \frac{N_i}{N}\bar{Y}_i$.

- Fact:
$$E(\hat{\mu}) = \mu,\ \text{Var}(\hat{\mu}) = \sum_{i=1}^K \left(\frac{N_i}{N}\right)^2 \frac{\sigma_i^2}{n_i}\frac{N_i - n_i}{N_i - 1}$$
- Fact: This variance is minimized when $n_i/n \propto N_i\sigma_i$.

Horvitz-Thompson (HT) estimator: Let $\tau = \sum_{i=1}^N y_i$ be the estimand. If we take a sample $S = \{Y_1, ..., Y_n\}$, the HT estimator is

$$\hat{\tau} = \sum_{i=1}^N \frac{I(y_i \in S)y_i}{P(y_i \in S)} = \sum_{j:y_j \in S} \frac{y_j}{P(y_j \in S)}$$

where I_i is an indicator of item y_i being sampled.

- Fact: $E(\hat{\tau}) = \tau$.
- Strategy: If we want to estimate μ, divide by N.
- Strategy: The first sum makes it easier to prove results; the second is easier to use in practice.
- Note: The HT estimator is unbiased, but it does not necessarily have low variance (Basu's elephants).

Problem 5

In simple one-stage cluster sampling, the original population is divided into (supposedly representative) clusters S_i, one cluster is sampled, and every item in the cluster is sampled. Suppose there are k clusters with N_i items in cluster i and N items total. We want to estimate the overall mean $\mu = \frac{1}{N}\sum_{j=1}^N y_j$. A cluster is sampled with a probability proportional to its size, everyone in the cluster (say, cluster i) is sampled, and the estimator $\hat{\mu} = \bar{y}_i = \frac{1}{N_i}\sum_{j:y_j \in S_i} y_j$ (the sample mean of the cluster) is used to estimate μ.

1. Is this the Horvitz Thompson estimator? If so, prove it. If not, find one.

Yes, it is the Horvitz Thompson estimator. The Horvitz Thompson estimator for the mean is $\frac{1}{N}\sum_{j=1}^N \frac{I_j y_j}{P(y_j \in S)}$. The probability that y_j will be sampled is the probability that y_j's cluster gets sampled: $P(I_j) = N_i/N$

where $y_j \in S_i$. Since all the other items in this cluster will be sampled and divided by the same probability N_i/N, we will have

$$\frac{1}{N}\sum_{j=1}^{N}\frac{I_j y_j}{P(y_j \in S)} = \frac{1}{N}\sum_{i=1}^{k}\frac{I_i \sum_{j:y_j \in S_i} y_j}{N_i/N} = \frac{1}{N}\sum_{i=1}^{k}\frac{I_i N_i \bar{y}_i}{N_i/N} = \sum_{i=1}^{k} I_i \bar{y}_i$$

Since only one of the I_i will be 1 (only one cluster is sampled), only a single $\bar{y}_i$ will be observed and used to estimate the mean.

2. Find the variance of this estimator.

Using the Horvitz Thompson notation,

$$\begin{aligned}
\text{Var}\left(\sum_{i=1}^{k} I_i \bar{y}_i\right) &= E\left(\left(\sum_{i=1}^{k} I_i \bar{y}_i\right)^2\right) - \left(E\left(\sum_{i=1}^{k} I_i \bar{y}_i\right)\right)^2 \\
&= E\left(\sum_{i=1}^{k}(I_i \bar{y}_i)^2\right) - \mu^2 \\
&= E\left(\sum_{i=1}^{k} I_i \bar{y}_i^2\right) - \mu^2 \\
&= \left(\sum_{i=1}^{k} P(I_i = 1)\bar{y}_i^2\right) - \mu^2 \\
&= \left(\frac{1}{N}\sum_{i=1}^{k} N_i \bar{y}_i^2\right) - \mu^2
\end{aligned}$$

The second equality came from the fact that $I_a I_b = 0$ for $a \neq b$.

3. Verify that the variance makes sense in the special cases of $N_1 = ... = N_k = 1$ (each item is its own cluster) and $\bar{y}_i = \mu$ (each cluster is perfectly representative of the overall mean).

When $N_1 = ... = N_k = 1$,

$$\left(\frac{1}{N}\sum_{i=1}^{k} N_i \bar{y}_i^2\right) - \mu^2 = \left(\frac{1}{N}\sum_{j=1}^{N} y_j^2\right) - \mu^2 = \sigma^2$$

which is the same variance we would have obtained from a SRS with only 1 observation as expected.

When $\bar{y}_i = \mu$,

$$\left(\frac{1}{N}\sum_{i=1}^{k} N_i \bar{y}_i^2\right) - \mu^2 = \left(\frac{1}{N}\sum_{i=1}^{k} N_i \mu^2\right) - \mu^2 = \mu^2 - \mu^2 = 0$$

which makes sense because we will always obtain the same mean μ regardless of which cluster we sample, so there will be no variance.

11: Resampling

Non-parametric bootstrap: Sample n values from the observed data $Y_1, ..., Y_n$ with replacement many times and calculate the estimator $\hat{\theta}_b^*$ for each sample. The bias of the estimator is estimated with $(\frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_b^*) - \hat{\theta}$, and the standard error is estimated with

$$\widehat{\text{SE}(\hat{\theta})} = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(\hat{\theta}_b^* - \overline{\hat{\theta}^*})^2}$$

Bootstrap confidence interval: We can use the non-parametric bootstrap to construct a confidence interval in the following ways:

1. **Normal approximation**: $\hat{\theta} \pm Q_{\mathcal{N}(0,1)}(1-\alpha/2)\widehat{\text{SE}(\hat{\theta})}$
2. **Percentile interval**: Use the empirical $\alpha/2$ and $1-\alpha/2$ quantiles of the bootstrapped θ_b^*
3. **Bootstrap t interval**: Approximate the distribution of $T = \frac{\hat{\theta}-\theta}{\widehat{\text{SE}(\hat{\theta})}}$ with

$$T^* = \frac{\hat{\theta}^* - \hat{\theta}}{\widehat{\text{SE}(\hat{\theta}^*)}}$$

Since $\widehat{\text{SE}(\hat{\theta}^*)}$ is usually unknown, we can run an additional layer of bootstrapping to estimate it. The bootstrapped interval is then $\left[\hat{\theta} - \hat{Q}^*(0.975)\widehat{\text{SE}(\hat{\theta})}, \hat{\theta} - \hat{Q}^*(0.025)\widehat{\text{SE}(\hat{\theta})}\right]$ where $\hat{Q}^*$ is the bootstrapped quantile of T^*.

Permutation testing: Suppose we have $X_1, ..., X_m \sim F_X$ and $Y_1, ..., Y_n \sim F_Y$ and we want to test $H_0 : F_X = F_Y$ vs $H_a : F_X \neq F_Y$. To run a permutation test, we choose a test statistic $T(\overrightarrow{X}, \overrightarrow{Y})$, compute the observed value of the test statistic t_0 from the data, permute the observations between groups keeping the sample sizes the same as originally and compute t^* for each many times, and find the proportion of times the t^* was as or more extreme than t_0. This proportion is the p-value.

- Note: T is often chosen to be the difference in sample means.

12: Causal inference

Assignment: The assignment W_j is 1 if subject j is in the treatment group and 0 otherwise.

Potential outcomes: $Y_j(w_1, ..., w_n)$.

- In words: The outcome for patient Y_j if the assignments were $w_1, ..., w_n$.

Treatment effect: $\tau_j = Y_j(w_1, ..., w_n) - Y_j(w_1', ..., w_n')$ is the effect of moving from one assignment to another.

Non-interference: The assignment of others has no effect on the potential outcomes of a particular subject: $Y_j(w_1, ..., w_n) = Y_j(w_j)$.

Assignment mechanism: $P(\overrightarrow{W} = \overrightarrow{w} | \overrightarrow{Y(0)}, \overrightarrow{Y(1)})$.

- In words: The joint PMF of assignments given the potential outcomes.
- Fact: In RCTs, $P(\overrightarrow{W} = \overrightarrow{w} | \overrightarrow{Y(0)}, \overrightarrow{Y(1)}) = P(\overrightarrow{W} = \overrightarrow{w})$; in observational studies, this is not necessarily the case. Note that this is not saying $W \perp\!\!\!\perp Y$.

Switching equation:

$$Y = Y(w) = Y(1)w + Y(0)(1-w)$$

Unconfoundedness: $\overrightarrow{W} \perp\!\!\!\perp (\overrightarrow{Y(0)}, \overrightarrow{Y(1)})$ given $\overrightarrow{X}$.

Population model: Assume $\{W_1, Y_1(0), Y_1(1)\}, ..., \{W_n, Y_n(0), Y_n(1)\}$ are independent (all three are random) and we condition on $\overrightarrow{W}$ in an RCT. The estimand is $E(\tau_1) = E(\tau_j)$ when the triples are i.i.d.

- Fact: If the triples are i.i.d. and the outcomes are binary, the MLE treatment effect given $\overrightarrow{w}$ is:

$$\widehat{E(\tau_1)} = \frac{\sum_{i=1}^n w_i Y_i}{\sum_{i=1}^n w_i} - \frac{\sum_{i=1}^n (1-w_i) Y_i}{\sum_{i=1}^n (1-w_i)}$$

This estimator is unbiased and has variance:

$$\operatorname{Var}\left(\widehat{E(\tau_1)}\right) = \frac{\theta_1(1-\theta_1)}{\sum_{i=1}^n w_i} - \frac{\theta_0(1-\theta_0)}{\sum_{i=1}^n (1-w_i)}$$

where $\theta_0 = P(Y_i = 1|W_i = 0)$ and $\theta_1 = P(Y_i = 1|W_i = 1)$.

Finite sample: In the finite sample setting, we treat the $y_j(0)$ and $y_j(1)$ as fixed, and the randomness comes from the W_j. Assume the assignments are independent of the potential outcomes. The estimand is $\hat{\tau} = \frac{1}{n}\sum_{i=1}^n (y_j(1) - y_j(0))$.

- Fact: The Method of Moments estimator is

$$\hat{\tau}_{\text{MOM}} = \frac{1}{n}\sum_{i=1}^n \left(\frac{W_jY_j}{E(W_j)} - \frac{(1-W_j)Y_j}{E(1-W_j)}\right)$$

This is unbiased and has conditional variance:

$$\operatorname{Var}(\hat{\tau}|\overrightarrow{Y(0)} = \overrightarrow{y(0)}, \overrightarrow{Y(1)} = \overrightarrow{y(1)}) = \frac{1}{n^2}\sum_{i=1}^n \left(\frac{y_i^2(1)}{E(W_i)} + \frac{y_i^2(0)}{E(1-W_i)} - (y_i(1) - y_i(0))^2\right)$$

Fisher null: For testing the finite sample treatment effect, $H_0 : \tau_j = 0$ for all j vs $H_a : \tau_j \neq 0$ for at least one j.

- Strategy: Fisher's null implies $Y_j = y_j(1) = y_j(0)$, so we can use a permutation test for $\hat{\tau}_{\text{MOM}}$, which we now call a **randomization test**.

Neyman's null: $H_0 : \bar{\tau} = 0$ vs. $H_a : \bar{\tau} \neq 0$.

Problem 6

In randomized control trials, it is sometimes the case that the treatment group does not actually take the treatment. Suppose that everyone assigned the non-treatment does not use the treatment, but suppose that each person assigned the treatment uses the treatment with probability p. In particular, let W_i be the indicator of whether the person actually took the treatment and T_i be the indicator of whether the person was assigned the treatment. Assume the $(W_i, Y_i(1), Y_i(0), T_i)$ quadruplets are i.i.d. across i and that the study is randomized. Also, assume that whether someone complies with the treatment is independent of the person's potential outcomes. The average treatment effect for the population is still $E(\tau_1) = E(Y_1(1) - Y_1(0))$.

1. Assuming we observe the T_i and Y_i but not the W_i, find the bias of the usual estimator

$$\widehat{E(\tau_1)} = \frac{\sum_{i=1}^n Y_i t_i}{\sum_{i=1}^n t_i} - \frac{\sum_{i=1}^n Y_i(1-t_i)}{\sum_{i=1}^n (1-t_i)}$$

Use this to make an unbiased estimator. Assume we know p. (We could estimate p from an in-depth follow-up study to check compliance in a subset of the individuals.)

In the estimator $\frac{\sum_{i=1}^n Y_i t_i}{\sum_{i=1}^n t_i} - \frac{\sum_{i=1}^n Y_i(1-t_i)}{\sum_{i=1}^n (1-t_i)}$, $Y_i(1-t_i)$ will still be $Y_i(0)$ only if t_i is 0 since w_i will be 0 and the switching equation still holds. Otherwise, it will be 0 because t_i will be 1. If $t_i = 1$, we must consider Y_i, which will be $Y_i(1)$ with probability p and $Y_i(0)$ with probability $1-p$. Thus, the expected value of the estimator is:

$$\begin{aligned}
E\left(\widehat{E(\tau_1)}\right) &= \frac{\sum_{i=1}^n E(Y_i|T_i=1)t_i}{\sum_{i=1}^n t_i} - E(Y_1(0)) \\
&= E(Y_1|T_1=1) - E(Y_1(0)) \\
&= E(Y_1|W_1=1,T_1=1)P(W_1=1|T_1=1) + E(Y_1|W_1=0,T_1=1)P(W_1=0|T_1=1) - E(Y_1(0)) \\
&= E(Y_1(1)|W_1=1,T_1=1)p + E(Y_1(0)|W_1=0,T_1=1)(1-p) - E(Y_1(0)) \\
&= E(Y_1(1))p + E(Y_1(0))(1-p) - E(Y_1(0)) \\
&= E(\tau_1)p
\end{aligned}$$

The bias is then $E(\tau_1)p - E(\tau_1)$. We could create an unbiased estimator by using

$$\frac{1}{p}\left(\frac{\sum_{i=1}^n Y_i t_i}{\sum_{i=1}^n t_i} - \frac{\sum_{i=1}^n Y_i(1-t_i)}{\sum_{i=1}^n (1-t_i)}\right)$$

2. Where did your derivation assume someone's compliance is independent of the person's potential outcomes? Why is this assumption important?

The step that said $E(Y_1(1)|W_1=1,T_1=1) = E(Y_1(1))$ assumed that $Y_1(1)$ was independent of both T_1 (because this is an RCT) and W_1 (because compliance is independent of the potential outcomes). This is important because it could be the case that people who are noncompliant are more likely to be sickly whether they use or don't use the treatment.

3. In a study conducted from 1993-1995 that evaluated the impact of insectiside treated nets (ITNs) on child mortality, there were 396 deaths per 16841.1 child-years in the groups that were provided ITNs and 461 deaths per 16494.8 child-years in the control groups. Compliance (proper usage of the nets) was assessed by surprise visits from a healthcare worker, and compliance varied over the course of the study, but assume the compliance was approximately 72%. What is the estimated effect of using ITNs on child mortality?

Using our unbiased estimator from above, we have

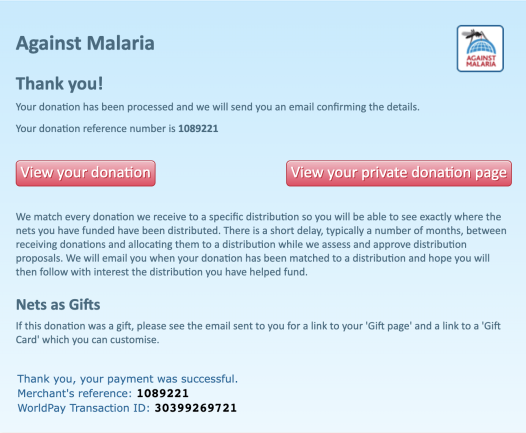
$$\frac{1}{p}\left(\frac{\sum_{i=1}^n Y_i t_i}{\sum_{i=1}^n t_i} - \frac{\sum_{i=1}^n Y_i(1-t_i)}{\sum_{i=1}^n (1-t_i)}\right) = \frac{1}{0.72}\left(\frac{396}{16841.1} - \frac{461}{16494.8}\right) = -0.0062$$

so the estimated effect is -0.0062 deaths per child-year or equivalently 6.2 deaths averted per 1000 child-years.

One last thing...

While this was primarily a final review, I've incorporated many examples from global health with the intention of showing that statistics isn't just about theory and tricks but that it can be used to figure out how to treat disease and save lives. A few of the examples today involved efforts to prevent malaria, a disease that kills an estimated 600,000 people annually, most of whom are children under the age of 5 in sub-Saharan Africa. Distribution and use of insecticide treated nets is one of the most well-studied ways of preventing malaria, and researchers who evaluate such interventions consider it one of the most cost-effective ways of preventing illness and saving lives of any current health intervention. After thorough independent evaluation, the Against Malaria Foundation has emerged as a non-profit extremely adept at these distributions, able to provide a net for \$2-\$5. Each net lasts 3-4 years, and for every 600 nets distributed, an estimated one child doesn't die, and 500-1000 cases of malaria are prevented.

While there's no pressure to do so, if anyone would like to go to againstmalaria.com and make a donation of any amount, I'll match up to the first \$1000 donated.



Against Malaria

Thank you!

Your donation has been processed and we will send you an email confirming the details.

Your donation reference number is **1089221**

We match every donation we receive to a specific distribution so you will be able to see exactly where the nets you have funded have been distributed. There is a short delay, typically a number of months, between receiving donations and allocating them to a distribution while we assess and approve distribution proposals. We will email you when your donation has been matched to a distribution and hope you will then follow with interest the distribution you have helped fund.

Nets as Gifts

If this donation was a gift, please see the email sent to you for a link to your 'Gift page' and a link to a 'Gift Card' which you can customise.

Thank you, your payment was successful.
Merchant's reference: **1089221**
WorldPay Transaction ID: **30399269721**

Thank you for coming today, and good luck on the final!