Chapter 7: Survey Sampling

Population Parameters

- N = population size
- All x_i are numerical values from the population
- Then population mean = $\frac{1}{N}\sum_{i=1}^{N} x_i$
- t = $\sum_{i=1}^{N} x_i = N\mu$
- Population variance: $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2$
- Population Confidence Interval for $X \sim N(\mu, \sigma^2)$
    - $P\left(\bar{X} - z(\alpha/2)\frac{\sigma^2}{\sqrt{n}} \leq \mu \leq \bar{X} + z(\alpha/2)\frac{\sigma^2}{\sqrt{n}}\right) = 1-\alpha$
    - eg: $P\left(\bar{X} - 2.58\frac{\sigma^2}{\sqrt{n}} \leq \mu \leq \bar{X} + 2.58\frac{\sigma^2}{\sqrt{n}}\right) = 0.99$
- CLT: $P\left(\frac{\bar{X}_n-\mu}{\sigma/\sqrt{n}} \leq z\right) \to \Phi(z) \quad \text{as } n \to \infty$

Sampling Parameters

- sample size n, values of the sample are X_i
- sample mean: $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. This is a random variable
- $$\operatorname{Var}\bar{X} = \frac{\sigma^2}{n}$$
- With random sampling: $$\operatorname{Var}(\bar{X}) = \frac{\sigma^2}{n}\left(1 - \frac{n-1}{N-1}\right)$$
- Without replacement: $\operatorname{Cov}(X_i, X_j) = -\frac{\sigma^2}{N-1}$

Sample CIs and Tests

- $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$
- Confidence interval for a normal sample mean: $P\left(-z(\alpha/2) \leq \frac{\bar{X}-\mu}{s/\sqrt{n}} \leq z(\alpha/2)\right) \approx 1-\alpha$
- $P\left(\bar{X} - z(\alpha/2)\frac{s}{\sqrt{n}} \leq \mu \leq \bar{X} + z(\alpha/2)\frac{s}{\sqrt{n}}\right) \approx 1-\alpha$
- Degrees of Freedom - number of observations minus the number of parameters used to explain the mean of all of those observations.

Chapter 8: Estimation of Parameters and Fitting Distributions

- Want some estimate of θ s.t. $\hat{\theta} = T(X_1, X_2, \ldots, X_n)$. The estimator is a function of the data alone.
- The observed data will be regarded as realizations of random variables $X_1, X_2, \ldots, X_n,$ whose joint distribution depends on an unknown parameter θ. An estimate of θ will be a function of $X_1, X_2, \ldots, X_n,$ and will hence be a random variable with a probability distribution called its sampling distribution. We use approximations of the sampling distribution to assess the variability of our parameter estimate, most often through standard error.
- There are three different kinds of θ in this setting. First there is θ the parameter which has a range of legal values. Then there is θ_0. When we need to single out the one true value of θ it is θ_0. In practice we don't know which value is true. Our MLE is $\hat{\theta}$.
- Bayesian estimation always treats a parameter as a random variable. Frequentist estimation sees the parameters as an unobserved value for which there is a true value.
- The set Ω of all possible values of a parameter θ or of a vector of parameters $(\theta_1, \ldots, \theta_k)$ is called the parameter space.
- Standard Error: the standard deviation of the estimate of a parameter. $\sigma_{\hat{\theta}} = \sqrt{\frac{\sigma^2}{n}}$. We generally do not know the true standard error, since it is a function of the true parameter.
- Estimated standard error - use the estimate of parameter: $s_{\hat{\theta}} = \sqrt{\frac{s^2}{n}}$
    - The estimated standard error is the standard deviation of the estimated parameter. Since we are often summing X_i's the above is really just the equation for $SD(\bar{X})$
- Unbiased: If $E(\hat{\lambda}) = \lambda$ then we say the estimate is unbiased.
- Consistent: An estimate $\hat{\theta}$ is said to be consistent if $P(|\hat{\theta}_n - \theta| > \epsilon) \to 0$ as $n \to \infty$ for all $\epsilon > 0$ and $\theta \in \Theta$

Fisher Information

- This measure has the intuitive properties that more data provide more information, and more precise data provide more information. The variance of the distribution tends to be inversely proportional to I.
- Define $\lambda(x|\theta) = \log f(x|\theta)$ - log likelihood function
- The Fisher information = $I(\theta) = E_\theta\left\{[\lambda'(X|\theta)]^2\right\} = -E_\theta\left[\lambda''(X|\theta)\right] = \mathrm{Var}_\theta[\lambda'(X|\theta)]$.
- Staring into it we see it is an expected squared slope of log likelihood. If the slope is large then small changes in θ change the log likelihood a lot. That should help separate likely from unlikely values. If that slope were zero then we get no effect of changing θ.
- $I_n(\theta) = nI(\theta)$ the Fisher information in a random sample of n observations is simply n times the Fisher information in a single observation.
- Can be used to compare sampling plans. Calculating the Fisher information for each and equating them will tell you something about the necessary parameters to yield the same information.

- Fisher information can be used to determine a lower bound for the variance of an arbitrary estimator of the parameter θ in a given problem -> Cramer - Rao

Method of Moments

- Generally have a sample of $X_i \sim iid$. We are only missing a parameter θ to know the distribution of these RVs
- kth sample moment defined as $\hat{\mu_k} = \frac{1}{n}\sum_{i=1}^{n} X_i^k$, where mu-hat is an estimate of the kth moment.
- If two parameters θ_1, θ_2 can be expressed in terms of the first two moments as $\theta_1 = f_1(\mu_1, \mu_2), \theta_2 = f_2(\mu_1, \mu_2)$ then the method of moments estimates are $\hat{\theta_1} = f_1(\hat{\mu_1}, \hat{\mu_2}), \hat{\theta_2} = f_2(\hat{\mu_1}, \hat{\mu_2})$
- Steps
    - Calculate low order moments, finding expressions for the moments in terms of the parameters. Usually need the same number of moments as parameters
    - Invert the expressions, finding parameters in terms of moments
    - Insert sample moments into the above expressions and you have your parameter estimates.
- MOM estimator for variance is different from unbiased s^2 : $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 = \hat{\mu}_2 - \hat{\mu}_1^2$
- MOM has a bias $\frac{\sigma_X^2}{2n} g''(\mu_X)$ for $\hat{\theta} = g(X)$
- When the standard error of an estimated parameter does not have a sampling distribution that can be obtained by plugging in parameters, we use bootstrap to estimate the SE.

Maximum Likelihood Estimation

- Random variables $X_1, X_2, \ldots, X_n$ have a joint density or frequency function $f(x_1, x_2, \ldots, x_n|\theta)$. Given observed values $X_i = x_i$, where $i = 1, \ldots, n$ the likelihood of θ as a function of $x_1, x_2, \ldots, x_n$ is defined as $lik(\theta) = f(x_1, x_2, \ldots, x_n|\theta)$. When the joint p.d.f. or the joint p.f. $fn(x|\theta)$ of the observations in a random sample is regarded as a function of θ forgiven values of $x_1, \ldots, x_n$, it is called the likelihood function.

- The MLE of θ is that value that maximizes the likelihood of f - it makes the observed data most probable. For large samples, MLE often yields a very good estimator. It is a value we believe the parameter to be near, but other estimates are likely to be better with smaller samples or any prior information. MLE is range respecting unlike MOM - we won't get an estimate that is beyond the domain of the parameter.

- If $\hat{\theta}$ is the maximum likelihood estimator of θ and if g is a one-to-one function, then $g(\hat{\theta})$ is the maximum likelihood estimator of $g(\theta)$. Or, if not one to one,, if we define $g(\theta)$ to be a function of θ, then $g(\hat{\theta})$ is an MLE of $g(\theta)$

- Therefore, the maximum likelihood estimate is the value of θ that assigned the highest probability to seeing the observed data. It is not necessarily the value of the parameter that appears to be most likely given the data.

- For iid X, $lik(\theta) = \prod_{i=1}^{n} f(X_i|\theta)$. Can also use the log likelihood: $l(\theta) = \prod_{i=1}^{n} log[f(X_i|\theta)]$
- Steps
    - Find joint density function, viewed as a function of the parameters.
    - Take partials with respect to the parameters (eg. μ, σ). Set partials to zero and solve for parameters.
    - Those parameters are the estimates - may need to solve system of equations to get RHS of

    equations just in terms of X's. May also need to use one estimated parameter to estimate the other (eg. for Normal, estimate μ and plug in this estimate into equation for σ estimate)
    - Ensure you have a maximum with 2nd derivative test and bound checking
- Range respecting: never gives illegal values (eg. Var < 0) unlike MOM. Transformations: $\hat{\theta} = MLE(\theta)$, then $g(\hat{\theta}) = MLE(g(\theta)$

Large Sample Theory for MLE

- The MLE from an iid sample is consistent (given smoothness of f)
- Define $I(\theta) = E\Big[\frac{\partial}{\partial\theta} logf(X|\theta)\Big]^2 = -E\Big[\frac{\partial^2}{\partial\theta^2} logf(X|\theta)\Big]$
- Under smoothness conditions on f ,the probability distribution of $\sqrt{nI(\theta_0)}(\hat{\theta} - \theta_0) \approx N(0,1)$. We say the MLE is asymptotically unbiased. We can interpret as $\hat{\theta} \approx N(\theta_0, \frac{1}{nI(\theta_0)})$
    - Note: does not hold if $f(X|\theta)$ support relies on θ, eg. $U[0,\theta]$
- Can also be interpreted as for an MLE from the log-likelihood function $l(\theta)$, the asymptotic variance is $\frac{1}{nI(\theta_0)} = -\frac{1}{El''(\theta_0)}$

CI from MLE

- 3 methods: exact. approximations with large samples, and bootstrap CI
- Exact example:
    - We have $\frac{(\bar{X}-\mu)}{s/\sqrt{n}} \sim t_{n-1}$
    - Then CI for μ = $P(\bar{X} \ + \frac{s}{\sqrt{n}} t_{n-1}(\alpha/2) \leq \mu \leq \bar{X} + \frac{s}{\sqrt{n}} t_{n-1}(1-\alpha/2)) = 1-\alpha$
    - Given $\frac{n\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-1}$ (or $\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$) CI for $\sigma = P(\frac{n\hat{\sigma}^2}{\chi^2_{n-1}(\alpha/2)} \leq \sigma^2 \leq \frac{n\hat{\sigma}^2}{\chi^2_{n-1}(1-\alpha/2)}) = 1-\alpha$. Note this is not symmetric about $\hat{\sigma}^2$
- Approximate
    - $\sqrt{nI(\hat{\theta})}(\hat{\theta} - \theta_0)$ tends to a standard normal distribution. Notice $\sqrt{nI(\hat{\theta})}$ is the theoretical limit SD from Cramer-Rao for efficient estimators - however it can work even if the MLE is not totally efficient.
    - There $P(-z(\alpha/2) \leq \sqrt{nI(\hat{\theta})}(\hat{\theta} - \theta_0) \leq z(\alpha/2)) \approx 1-\alpha$
    - This gives us $\hat{\theta} \pm z(\alpha/2)\frac{1}{\sqrt{nI(\hat{\theta})}}$. Notice this is z-stat times SE, just like any confidence interval.
    - For multinomial: $Var(\hat{\theta}) \approx \frac{1}{El'(\theta_0)^2} = -\frac{1}{El''(\theta_0)}$ and the MLE is approximately normally distributed, then this is used for SE in CI
- Bootstrap
    - Take estimate $\hat{\theta}$, generate B samples from distribution with parameter $\hat{\theta}$, construct estimate of parameter θ^*. Then the distribution of $\hat{\theta} - \theta_0 \approx \theta^* - \hat{\theta}$. Subtract $\hat{\theta}$ from each $\theta^*_j, \ j \in (1,2,\ldots,B)$. Define $\alpha/2$ and $1-\alpha/2$ as quantiles of this distribution by $\underline{\delta}, \ \bar{\delta}$.
    - Confidence interval is then $(\hat{\theta} - \bar{\delta}, \hat{\theta} - \underline{\delta})$

Cramer-Rao Lower Bounds / Efficiency

- Given two estimates $\hat{\theta}, \bar{\theta}$ of the same parameter, the efficiency of $\hat{\theta}$ relative to $\bar{\theta}$ is defined to be $eff(\theta,\hat{}\bar{\theta}) = \frac{Var(\bar{\theta})}{Var(\hat{\theta})}$

- If eff is smaller than 1, then theta-hat has a larger variance than theta-bar. Most meaningful when the estimates have the same bias. if Var(estimator) = $\frac{c}{n}$, then the efficiency is the ratio os sample sizes necessary to obtain the same variance for both estimators.
- Cramer Rao inequality: $X_1, X_2, \ldots, X_n$ iid with density function $f(x|\theta)$. Let $T = t(X_1, X_2, \ldots, X_n)$ be an unbiased estimate of θ. Then $Var(T) \geq \frac{1}{nI(\theta)}$. If it is an equality, then T is an efficient estimator.
  - More general version: T statistic with finite variance, $m(\theta) = E_\theta(T)$, $\operatorname{Var}_\theta(T) \geq \frac{[m'(\theta)]^2}{nI(\theta)}$. When T goes to theta in expectation, then the numerator is the derivative of theta = 1.
- The variance of an unbiased estimator of θ cannot be smaller than the reciprocal of the Fisher information in the sample.
- Theorem A gives a lower bound on the variance of any unbiased estimate. An unbiased estimate whose variance achieves this lower bound is said to be efficient. Since the asymptotic variance of a maximum likelihood estimate is equal to the lower bound, maximum likelihood estimates are said to be asymptotically efficient. Ie. MLE is a good estimator for large enough n, though could be beaten by some Bayesian estimators with bias.
- If T is an efficient estimator of $m(\theta)$, then among all unbiased estimators of $m(\theta)$, T will have the smallest variance for every possible value of θ

Bayesian Estimation

- The parameter of interest θ is actually the observed value of a random variable Θ, eg. we might think a failure rate is modeled by an exponential process. When one treats the parameter as a random variable, the name "prior distribution" is merely another name for the marginal distribution of the parameter. Assume Θ is a continuous RV.
- We call $f_\Theta(\theta)$ the prior distribution of Θ and $f_{X|\Theta}(x|\theta)$ the posterior distribution of Θ. Therefore we can get the joint distribution of X and Θ by $f_{X,\Theta}(x,\theta) = f_{X|\Theta}(x|\theta)f_\Theta(\theta)$ When one treats the parameter as a random variable, the name "posterior distribution" is merely another name for the conditional distribution of the parameter given the data. For many random variables iid (ie data), $f_n(x_1, \ldots, x_n|\theta) = f(x_1|\theta)\ldots f(x_n|\theta)$.
- The posterior is proportional to likelihood x prior : $f_{\Theta|X}(\theta|x) \propto f_{X|\Theta}(x|\theta)f_\Theta(\theta)$.
  - Why? The $\int f_{X|\Theta}(x|\theta)f_\Theta(\theta)\, d\theta$ term (ie. the marginal of X) is simply the integral of the numerator over all possible values of θ. Although the value of this integral depends on the observed values $x_1, \ldots, x_n$, it does not depend on θ and it may be treated as a constant when $\frac{f_{X,\Theta}}{f_x(x)}$ is regarded as a p.d.f. of θ.
  - The appropriate constant factor that will establish the equality of the two sides Bayes rule can be determined at any time by using the fact that $\int_\Omega f(\theta|x)d\theta = 1$, because $f(\theta|x)$ is a p.d.f. of θ. Often we can do this without integration by recognizing the posterior as a known probability distribution missing a constant.
- Essentially $f_{\Theta|X}(\theta|x) = \frac{f_{X,\Theta}}{f_x(x)} = \frac{f_{X|\Theta}(x|\theta)f_\Theta(\theta)}{\int f_{X|\Theta}(x|\theta)f_\Theta(\theta)\, d\theta}$. The marginal of X is the joint pdf integrated over all values of θ.
- Steps to find the posterior distribution
  - Take a prior distribution, eg. $\Theta \sim U(a,b)$
  - Find $f(x|\theta)$, ie. the distribution of the data in terms of the parameter. Eg. $X \sim N(\theta, \sigma^2)$

    - Calculate $f(x|\theta)f(\theta)$
    - Try to figure out constant from a recognized probability distribution. Last resort integrate the denominator of Bayes, ie $\int f_{X|\Theta}(x|\theta)f_\Theta(\theta)\,d\theta$
- We can also calculate sequentially using one data point at a time to obtain the same posterior, using the posterior from one observation as the prior for the next. For improper priors, we use some constant, and then $posterior(\theta) \propto likelihood(\theta) \times constant$
- Steps for finding an estimator $\hat{\theta}$
    - It could be the mean, median or mode of the posterior distribution.
    - The posterior mean minimizes the posterior mean squared error $\mathbb{E}_{\Theta|X}\left((\hat{\theta}-\Theta)^2|x\right)$
    - The posterior median minimizes the posterior mean absolute error $\mathbb{E}_{\Theta|X}\left(|\hat{\theta}-\Theta||x\right)$
    - Variance of the estimate is taken to be the variance of the posterior distribution.
    - Analogue of CI is taking the $\alpha/2, (1-\alpha/2)$ percentiles of the distribution.
    - Alternatively, high posterior density interval (HPD) - place horizontal line on posterior density and move it down until the area cut by the line contains $1-\alpha$ of the density area. Will only be different from percentile CI if not symmetrical.
- Bayes estimators are consistent
- The Bayesian interval is a probability statement referring to the state of knowledge about θ given the observed data, regarding θ as a random variable. The frequentist confidence interval is based on a probability statement about the possible values of the observations, regarding θ as a constant, albeit unknown.
- The posterior distribution is approximately normal with the mean equal the the maximum likelihood estimate, $\hat{\theta}$, and variance approximately equal to $-\left[l''(\hat{\theta})\right]^{-1}$

Gibbs Sampling

- Instead of relying on analytically finding normalizing constant or using conjugates, can computationally find what the constant needs to be using Monte Carlo. We have a joint density that is hard to estimate. Suppose, though, that we can easily sample from the conditional distributions p(x|y) and p(y|x).
- Gibbs Sampling alternates back and forth between the two conditional distributions:

1. Choose starting value for θ_0, usually $\bar{x}$
2. Sample ξ_0 from distribution given θ_0
3. Repeat with distribution $p(\theta_0|\Xi=\xi_0)$
4. Repeat incremented data values by 1, alternating between conditional distributions

- Some burn in required before simulations have vale. $(\xi_k, \theta_k), k = 1, \ldots, N$ approximates pull from the posterior.

Conjugates

- In science, flat / uninformative priors are often used because another prior may determine the shape of the posterior. Improper priors do not integrate to one, so the posterior may not integrate to one either.
- Make the calculations in Bayesian estimation easier by telling you what distribution the posterior should be

- For each of the most popular statistical models, there exists a family of distributions for the parameter with a very special property. If the prior distribution is chosen to be a member of that family, then the posterior distribution will also be a member of that family. Such a family of distributions is called a conjugate family.
- Prior / Posterior - Likelihood
  - Beta - Bernoulli
  - Beta - Binomial
  - Beta - Geometric
  - Gamma - Poisson
  - Normal - Normal (precision or variance)
  - Gamma - Exponential
  - Gamma - Gamma

Sufficiency

- Given the initial conditions for Cramer Rao above.
- Imagine one statistician who can observe the data and one who only gets a statistic about the data. Is there statistic such that observing the data provides no additional information. A sufficient statistic is sufficient for being able to compute the likelihood function, and hence it is sufficient for performing any inference that depends on the data only through the likelihood function. M.L.E.'s and anything based on posterior distributions depend on the data only through the likelihood function.
- The concept of sufficiency arises as an attempt to answer the following question: Is there a statistic, a function $T(X_1, X_2, \ldots, X_n)$, that contains all the information in the sample about θ? If so, a reduction of the original data to this statistic without loss of information is possible. For example, in series of Bernoulli trials, total number of successes contains all of the information about p that exists in the sample.
- Sufficiency: A statistic $T(X_1, X_2, \ldots, X_n)$ is said to be sufficient for θ if the conditional distribution of $X_1, X_2, \ldots, X_n$ given $T = t$, does not depend on θ for any value of t.
- In other words, given the value of T , which is called a sufficient statistic, we can gain no more knowledge about θ from knowing more about the probability distribution of $X_1, X_2, \ldots, X_n$. An observable random variable
- Factorization: A necessary and sufficient condition for $T(X_1, X_2, \ldots, X_n)$ to be sufficient for a parameter θ is that the joint probability function (density function or frequency function) factors in the form: $f(x_1, \ldots, x_n|\theta) = g[T(x_1, \ldots, x_n), \theta]h(x_1, \ldots, x_n)$. u depends on x but not theta. g depends on theta but only depends on x through the statistic.
- The MLE is found through maximizing $g[T(x_1, \ldots, x_n), \theta]$. If T is sufficient for θ, the MLE is a function of T

Rao Blackwell Theorem

- Let $\hat{\theta}$ be an estimator of θ with existing expectation for all theta. If T is sufficient for theta and $\bar{\theta} = E(\hat{\theta}|T)$. Then for all theta: $E(\bar{\theta} - \theta) \leq E(\hat{\theta} - \theta)^2$
- Rationale for basing estimators on sufficient stats if they exist. Conditioning on T is sure to give a function of the data, not a function of the true parameter theta that cannot be observed.