1 Introduction and Recap

- $E(aX + bY) = aE(X) + bE(Y)$
- $Var(aX + bY) = a^2Var(X) + b^2Var(Y) + 2abCov(X, Y)$
- $Cov(X, Y) = E[(X - E(X))(Y - E(Y))] = E(XY) - E(X)E(Y)$
- $Cov(aX + b, cY + d) = abCov(X, Y)$
- Law of Large Numbers: As $n \to \infty$, $\bar{X} \to \mu$
- Central Limit Theorem: As $n \to \infty$, distribution of sum S_n when standardised converges to Z

2 Distributions, Normal and Related

Normal: $X \sim N(\mu, \sigma^2)$

- $f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

Chi-square: $U \sim \chi^2_1$ where $U = Z^2$

- $E(U) = 1$
- $Var(U) = 2$
- χ^2_1 is Gamma with $\alpha = \frac{1}{2}$, $\lambda = \frac{1}{2}$

Chi-square n: $V \sim \chi^2_n$ where $V = U_1 + \ldots + U_n$

- $E(V) = n$
- $Var(V) = 2n$
- χ^2_n is Gamma with $\alpha = \frac{n}{2}$, $\lambda = \frac{1}{2}$
- As $n \to \infty$, V is approximately normal by CLT
- If $X \sim \chi^2_m$ and $Y \sim \chi^2_n$ are independent, then $X + Y \sim \chi^2_{m+n}$

Gamma distribution: $X \sim Gamma(\alpha, \lambda)$

- $g(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}$, for $x \geq 0$
- α is the *shape* parameter
- λ is the *rate* parameter — how fast it 'dies out'
- If $X \sim Gamma(\alpha_1, \lambda)$ and $Y \sim Gamma(\alpha_2, \lambda)$ are independent, then $X + Y \sim Gamma(\alpha_1 + \alpha_2, \lambda)$

t distribution: $t_n = \frac{Z}{\sqrt{V_n/n}}$, where $Z \sim N(0, 1)$ and $V_n \sim \chi^2_n$

- t_n has a t distribution with n degrees of freedom
- $f(t) = \frac{\Gamma((n+1)/2)}{\sqrt{n\pi}\Gamma(n/2)} (1 + \frac{t^2}{n})^{-(n+1)/2}$
- As $n \to \infty$, t_n converges to Z, since V_n/n converges to its mean 1

F distribution: $W = \frac{U/m}{V/n}$, where $U \sim \chi^2_m$ and $V \sim \chi^2_n$

- W has an F distribution with (m, n) degrees of freedom
- $E(W) = \frac{n}{n-2}$ for $n > 2$

- If $X \sim t_n$, then X has distribution of $\frac{Z}{\sqrt{V_n/n}}$ and X^2 has distribution of $\frac{Z^2/1}{V_n/n}$ (F distribution with $(1, n)$ degrees of freedom)
- As $n \to \infty$, $W \approx U/m$ since $\frac{V}{n} \to 1$

Binomial: $X \sim Bin(n, p)$

- $E(X) = np$
- $Var(X) = np(1-p)$

Poisson: $X \sim Po(\lambda)$

- $P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$
- $E(X) = \lambda$
- $Var(X) = \lambda$

Exponential: $X \sim Exp(\lambda)$, $\lambda > 0$

- $f(x) = \lambda e^{-\lambda x}$
- $E(X) = \frac{1}{\lambda}$
- $Var(X) = \frac{1}{\lambda^2}$
- $Exp(\lambda)$ is Gamma with $\alpha = 1$, same λ

Multinomial: $\boldsymbol{X} \sim Multinomial(\boldsymbol{p})$

- $f(x_1, \ldots, x_r) = \binom{n}{x_1 \ldots x_r} \prod_{i=1}^r p_i^{x_i}$
- $\ell(\mathbf{p}) = \kappa + \sum_{i=1}^r x_i \log p_i$, where $\kappa = \log \binom{n}{x_1 \ldots x_r}$ does not depend on $\mathbf{p}$

3 Survey Samping A: SRS, Estimation

3.1 Summary of Key Statistics

Population

- $\mu = \frac{1}{N} \sum_i x_i$
- $\tau = \sum_i x_i$
- $\sigma^2 = \frac{1}{N} \sum_i (x_i - \mu)^2 = \frac{1}{N} \sum_i x_i^2 - \mu^2 = \mu_2^2 - \mu_1^2$

Sample

- $\bar{x} = \frac{1}{n} \sum_i x_i$
- $\hat{\sigma}^2 = \frac{1}{n} \sum_i (x_i - \bar{x})^2 = \frac{1}{n} \sum_i x_i^2 - \bar{x}^2$
- $s^2 = \frac{n}{n-1} \hat{\sigma}^2 = \frac{1}{n-1} \sum_i x_i^2 - \frac{n}{n-1} \bar{x}^2$
- $s = \sqrt{\frac{\sum_i x_i^2 - n\bar{x}^2}{n-1}} = \sqrt{\frac{\sum_i x_i^2 - \frac{1}{n}\sum_i x_i}{n-1}}$

Finite population correction factor: $\frac{N-n}{N-1} = 1 - \frac{n-1}{N-1}$

- Used when drawing without replacement
- Close to 1 when sampling fraction $\frac{n}{N}$ is small

3.2 Simple Random Sampling

SRS: Sampling *without replacement*

- $E(\bar{X}) = \mu$
- $Var(\bar{X}) = \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1}$ (SRS)

Special Case: Proportion of 0/1

$\bar{X} = \hat{p}$: sample mean i.e. the proportion of 1s

- $E(\hat{p}) = p$
- $Var(\hat{p}) = \frac{p(1-p)}{n} \cdot \frac{N-n}{N-1}$ (SRS)

3.3 Estimation

Standard error: SE of estimate = SD of estimator

- Hence SE of $\bar{x}$ = SD of $\bar{X} = \frac{\sigma}{\sqrt{n}} \approx \frac{s}{\sqrt{n}}$

Biased and unbiased estimators

- $\hat{\sigma}^2$ is biased estimator of σ^2
- $s^2 = \frac{n}{n-1}\hat{\sigma}^2$ is unbiased estimator of σ^2

Summary

Parameter	Estimate	SE	Estimated SE
μ	$\bar{x}$	$\frac{\sigma}{\sqrt{n}}$	$\frac{s}{\sqrt{n}}$
p	$\hat{p}$	$\sqrt{\frac{p(1-p)}{n}}$	$\sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}}$

4 Survey Sampling B: Confidence Intervals, Measurement Model

Normal approximation for $\bar{X}$

- $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ is approximately $N(0,1)$ by CLT for large n

4.1 Confidence Intervals for $\bar{X}$ (Normal)

Let α be our significance level, and we want to construct $(1-\alpha)$ CI for μ.

$(1-\alpha)$ CI for μ: $(\bar{x} - z_{\alpha/2}\frac{s}{\sqrt{n}}, \bar{x} + z_{\alpha/2}\frac{s}{\sqrt{n}})$

Interpretation: probability of a constructed CI containing μ is $(1-\alpha)$

5 Survey Sampling C: Ratio Estimation

Setup: each member of the population has two characteristics, i.e. $(x_1, y_1), \ldots, (x_N, y_N)$

Ratio parameter: $r = \frac{\mu_Y}{\mu_X}$

Estimator of r: $R = \frac{\bar{Y}}{\bar{X}}$

- Is this a good or bad estimator? Bias? Variance?
- Assume that $\mu_x \approx \bar{X}$ and $\mu_y \approx \bar{Y}$ with large n
- $E(R) \approx r + \frac{1}{n}(\frac{N-n}{N-1})\frac{1}{\mu_x}(r\sigma_x^2 - \rho\sigma_x\sigma_y)$

- $E(R) \approx R + \frac{1}{n}(\frac{N-n}{N-1})\frac{1}{\bar{x}}(Rs_x^2 - \rho s_x s_y)$ (estimate)
- $Var(R) \approx \frac{1}{n}(\frac{N-n}{N-1})\frac{1}{\mu_x^2}(r^2\sigma_x^2 + \sigma_y^2 - 2r\sigma_{xy})$
- $s_R^2 \approx \frac{1}{n}(\frac{N-n}{N-1})\frac{1}{\bar{x}^2}(R^2 s_x^2 + s_y^2 - 2Rs_{xy})$ (estimate)
- $\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$ — between -1 and 1, unitless

Estimations

- Estimate σ_x^2 using $s_x^2 = \frac{1}{n-1}\sum_i (x_i - \bar{x})^2$
- Estimate σ_y^2 using $s_y^2 = \frac{1}{n-1}\sum_i (y_i - \bar{y})^2$
- Estimate σ_{xy} using $s_{xy} = \frac{1}{n-1}\sum_i (x_i - \bar{x})(y_i - \bar{y})$
- Estimate ρ using $\frac{s_{xy}}{s_x s_y}$

Confidence interval for r

...

6 Parameter Estimation A: Method of Moments

6.1 General Estimation Problem

For an unknown parameter θ, we wish to estimate it using estimator $\hat{\theta}$, using realisations $x_1 \dots x_n$.

- $Bias = E(\hat{\theta}) - \theta$
- $SE = SD(\hat{\theta})$
- $MSE = E(\hat{\theta} - \theta)^2 = SE^2 + Bias^2$

$$\begin{aligned} MSE &= E(\hat{\theta} - \theta)^2 \\ &= E\left((\hat{\theta} - E(\hat{\theta})) - (\theta - E(\hat{\theta}))\right)^2 \\ &= E\left(\hat{\theta} - E(\hat{\theta})\right)^2 + \left(\theta - E(\hat{\theta})\right)^2 - 0 \\ &= SE^2 + Bias^2 \end{aligned}$$

6.2 Consistent Estimators

Estimator $\hat{\theta}$ is *consistent* if as $n \to \infty$, $\hat{\theta} \to \theta$, i.e. it converges towards what it's trying to estimate as sample size increases

6.3 Method of Moments

MOM estimators may be biased or unbiased, but are generally consistent

- $\mu_k = E(X^k)$, i.e. k-th moment of X
- Estimated by $\hat{\mu}_k = \frac{1}{n}\sum_{i=1}^n x_i^k$

MOM:

1. Express μ_k in terms of parameters θ: can use common knowledge, density function, etc.
2. Rearrange the equations to express θ in terms of μ_k.
3. Finally, you can express MOM estimator $\hat{\theta}$ in terms of $\hat{\mu}_k$, by substituting μ_k with its estimator $\hat{\mu}_k$.

6.4 Example: Poisson

Let $X_i \sim Po(\lambda)$, $\theta = \lambda$

- Step 1: we know that $\mu_1 = \lambda$
- Step 2: we rearrange to get $\lambda = \mu_1$
- Step 3: we estimate to get $\hat{\lambda} = \hat{\mu}_1 = \bar{X}$

6.5 Example: Normal

Let $X_i \sim N(\mu, \sigma^2)$, $\theta = (\mu, \sigma^2)$

- Step 1: we know that $\mu_1 = \mu$ and $\mu_2 = \sigma^2 + \mu^2$
- Step 2: we rearrange to get $\mu = \mu_1$, and $\sigma^2 = \mu_2 - \mu_1^2$
- Step 3: we estimate to get $\hat{\mu} = \hat{\mu}_1 = \bar{X}$, and $\hat{\sigma}^2 = \frac{1}{n}\sum_i X_i^2 - (\bar{X})^2 = \frac{1}{n}\sum_i (X_i - \bar{X})^2$ (last step why?)

6.6 Example: Gamma

Let $X_i \sim Gamma(\lambda, \alpha)$, $\theta = (\lambda, \alpha)$, $f(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}$

- Step 1: we know that $\mu_1 = \frac{\alpha}{\lambda}$ and $\mu_2 = \frac{\alpha(\alpha+1)}{\lambda^2}$
- Step 2: we rearrange to get $\lambda = \frac{\mu_1}{\mu_2 - \mu_1^2}$, and $\alpha = \frac{\mu_1^2}{\mu_2 - \mu_1^2}$
- Step 3: we estimate to get $\hat{\lambda} = \frac{\hat{\mu}_1}{\hat{\mu}_2 - \hat{\mu}_1^2} = \frac{\bar{X}}{\hat{\sigma}^2}$, $\hat{\alpha} = \frac{\hat{\mu}_1^2}{\hat{\mu}_2 - \hat{\mu}_1^2} = \frac{\bar{X}^2}{\hat{\sigma}^2}$

6.7 Example: Angular Distribution

Let X be such that $f(x) = \frac{1+\alpha x}{2}$ where $x \in [-1, 1]$ and $\alpha \in [-1, 1]$ is unknown parameter

- Step 1: we derive $\mu_1 = \int_{-1}^{1} x \cdot \frac{1+\alpha x}{2} = \frac{\alpha}{3}$
- Step 2: we rearrange to get $\alpha = 3\mu_1$
- Step 3: we estimate to get $\hat{\alpha} = 3\hat{\mu}_1 = 3\bar{X}$

7 Parameter Estimation B: Bootstrap and Monte Carlo

Bootstrap approximation: suppose estimate of $\hat{\theta}$ is actual value of θ.

Monte Carlo approximation: use a large sample to approximate an expectation or SD.

7.1 Bootstrap and Monte Carlo

Purpose: approximate the *bias* and *SE* of our estimates, when there's no closed form way to approximate it

- Suppose we have our estimate of θ using $\hat{\theta} = 1.67$
- Bootstrap approximation:
  - $Bias(1.67) = E_\theta(\hat{\theta}) - \theta \approx E_{1.67}(\hat{1.67}) - 1.67$
  - $SE(1.67) = SD_\theta(\hat{\theta}) \approx SD_{1.67}(\hat{1.67})$
- Generate (perhaps) 10,000 realisations for $\hat{1.67}$, each time using n samples
- Monte Carlo approximation:
  - $E_{1.67}(\hat{1.67}) - 1.67 \approx 0.09$

- $SD_{1.67}(\hat{1.67}) \approx 0.35$

- Hence we estimate θ to be $1.67 - 0.09 \pm 0.35$

- To recap:
  - $Bias(1.67) = E(\hat{\theta}) - \theta \approx E(\hat{1.67}) - 1.67$ (by Bootstrap) ≈ 0.09 (by Monte Carlo)
  - $SE(1.67) = SD(\hat{\theta}) \approx SD(\hat{1.67})$ (by Bootstrap) ≈ 0.35 (by Monte Carlo)

8 Parameter Estimation C: Maximum Likelihood

ML similarly gives consistent estimators, and have the smallest SE among all consistent estimators (asymptotically the most efficient)

Likelihood function: likelihood of θ is given as a function of the given data

- Likelihood function: $L(\theta) = \prod_{i=1}^{n} f(x_i|\theta)$
- Loglikelihood function: $\ell(\theta) = \sum_{i=1}^{n} \log f(x_i|\theta)$
- Random likelihood function: $L(\theta) = \prod_{i=1}^{n} f(X_i|\theta)$
- Random loglikelihood function: $\ell(\theta) = \sum_{i=1}^{n} \log f(X_i|\theta)$

8.1 ML method

ML estimator $\hat{\theta}_0$ is the value that maximises the likelihood function $L(\theta)$, or equivalently the loglikelihood function $\ell(\theta)$.

One way to find $\hat{\theta}$ is to set $\ell(\hat{\theta}) = 0$, and confirm $\ell''(\theta) < 0$ at that value of $\hat{\theta}$. But alternatively, there are other ways too (e.g. if $L(\theta) = \theta(1-\theta)$, then obviously $\hat{\theta} = \frac{1}{2}$, no need to find loglikelihood nor differentiate).

ML:

1. Find the loglikelihood function $\ell(\theta)$
2. Find its derivatives $\ell'(\theta)$ and $\ell''(\theta)$
3. Maximize the loglikelihood function by setting $\ell'(\hat{\theta}) = 0$ and deriving the value of $\hat{\theta}$
4. Confirm it is a maximum with $\ell''(\hat{\theta}) < 0$

Tip: with loglikelihood function $\ell(\theta)$, you don't always have to find a fully closed form expression for it. It can be something like $\ell(\theta) = \kappa + \log \frac{x_1}{\theta} + \log \frac{x_2}{1-\theta}$, and the constant will go away when differentiating.

Note: if your θ is a vector, e.g. $\theta = (\mu, \sigma^2)$, then the derivatives have to be of form $\frac{d\ell}{d\mu}(\hat{\mu}, \hat{\sigma}^2)$ and $\frac{d\ell}{d\sigma^2}(\hat{\mu}, \hat{\sigma}^2)$

8.2 Confidence Intervals based on MLE

If n is large, then MLEs are asymptotically normal (by CLT), so we can construct approximate CIs

Normal Case

In the normal case, we can have *exact* CIs

- MLE for μ is $\bar{X}$
- MLE for σ is $\hat{\sigma}^2$
- CI for μ is $(\bar{x} - t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}, \bar{x} + t_{n-1,\alpha/2}\frac{s}{\sqrt{n}})$
- CI for σ^2 is $(\frac{n\hat{\sigma}^2}{\chi^2_{n-1,\alpha/2}}, \frac{n\hat{\sigma}^2}{\chi^2_{n-1,1-\alpha/2}})$, since $\frac{n\hat{\sigma^2}}{\sigma^2} \sim \chi^2_{n-1}$

How to find exact SE of $\bar{X}$ estimate?

- $Var(\bar{X}) = \frac{\sigma^2}{n}$
- $SD(\bar{X}) = \frac{\sigma}{\sqrt{n}}$
- Then can approximate σ with s (or $\hat{\sigma}$) using bootstrap

How to find exact SE of $\hat{\sigma}^2$ estimate?

- $Var(\hat{\sigma}^2) = \frac{\sigma^4}{n^2} Var(\frac{n\hat{\sigma}^2}{\sigma^2}) = \frac{\sigma^4}{n^2} \cdot 2(n-1)$
- $SD(\hat{\sigma}^2) = \frac{\sigma^2}{n} \cdot \sqrt{2(n-1)}$
- Then can approximate σ^2 with $\hat{\sigma}^2$ using bootstrap

9 Fisher Information

Fisher information matrix: $p \times p$ matrix, where $\theta \in \Theta \subset \mathbb{R}^p$

- $I(\theta) = -\int_{-\infty}^{\infty} \left[\frac{\partial}{\partial \theta^2} \log f(x|\theta)\right] f(x|\theta) dx$
- $I_{ij}(\theta) = -\int_{-\infty}^{\infty} \left[\frac{\partial^2}{\partial \theta_i \partial \theta_j} \log f(x|\theta)\right] f(x|\theta)\ dx$
- $I(\theta) = -E\left[\frac{\partial^2}{\partial \theta^2} \log f(X|\theta)\right]$, where X has density $f(x|\theta)$

Interpretation: $I(\theta)$ indicates amount of information about θ in *one* sample of $X \sim f(x|\theta)$

- If you have n independent samples, then the amount of information is just $nI(\theta)$ (by linearity of expectation)

How to find Fisher Information (note: one sample of X only!)

1. Find random logdensity $\log f(X|\theta)$
2. Differentiate w.r.t. θ to get $\frac{\partial}{\partial \theta} \log f(X|\theta)$
3. Differentiate w.r.t. θ to get $\frac{\partial^2}{\partial \theta^2} \log f(X|\theta)$
4. Take negative expectation to get $I(\theta) = -E\left[\frac{\partial^2}{\partial \theta^2} \log f(X|\theta)\right]$

Fisher Information and Variance

Distribution	Parameter	MLE	Variance
$Po(\lambda)$	λ	X	λ
$Ber(p)$	p	X	$p(1-p)$
$Bin(n,p)$	p	$\frac{X}{n}$	$\frac{p(1-p)}{n}$
$HWE\ Trinom$	θ	$\frac{X_2+2X_3}{2n}$	$\frac{\theta(1-\theta)}{2n}$
$General\ Trinom$	(p_1,p_2)	$(\frac{X_1}{n}, \frac{X_2}{n})$	$\frac{1}{n}\begin{bmatrix} p_1(1-p_1) & -p_1p_2 \\ -p_1p_2 & p_2(1-p_2) \end{bmatrix}$

In these cases, sample size = 1. Note that in these cases, $Var(\hat{\theta}) = I(\theta)^{-1}$ — the larger the information, the smaller the variance. (But in general, this equality is not true)

($\star$) Recall that $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \Leftrightarrow A^{-1} = \frac{1}{ad-bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$

10 Large Sample Theory for MLE

MLEs are *consistent*, i.e. converges to what it's trying to estimate as $n \to \infty$

Asymptotic normality of MLE: $\hat{\theta} \sim N\left(\theta, \frac{I(\theta)^{-1}}{n}\right)$ approxiately for large n

Approximate SE: $SE(\hat{\theta})$ (estimate) $= SD(\hat{\theta}) \approx \sqrt{\frac{I(\theta)^{-1}}{n}}$ (for large n) $\approx \sqrt{\frac{I(\hat{\theta})^{-1}}{n}}$ (bootstrap)

Approximate $(1-\alpha)$ CI: $\left(\hat{\theta} - z_{\alpha/2}\sqrt{\frac{I(\hat{\theta})^{-1}}{n}}, \hat{\theta} + z_{\alpha/2}\sqrt{\frac{I(\hat{\theta})^{-1}}{n}}\right)$ for large n

11 Efficiency

We only ever talk about efficiency (how good an estimator is) for *unbiased* estimators. It makes no sense to talk about efficiency for biased estimators in general.

11.1 Cramer-Rao Lower Bound

Theorem: If $\hat{\theta}$ is *unbiased*, then for every $\theta \in \Theta$, $Var(\hat{\theta}) \geq \frac{I(\theta)^{-1}}{n}$

Cramer-Rao lower bound (CRLB): $\frac{I(\theta)^{-1}}{n}$ is the best (i.e. lowest) variance you can ask from any unbiased estimator. No unbiased estimator can do better than this.

11.2 Efficiency

Efficient: $\hat{\theta}$ is *efficient* if $Var(\hat{\theta}) = \frac{I(\theta)^{-1}}{n}$ for every $\theta \in \Theta$

Efficiency: $Eff(\hat{\theta}) = \frac{I(\theta)^{-1}/n}{Var(\hat{\theta})}$. Efficiency is always ≤ 1.

Efficiency of ML estimators: By asymptotic normality theorem, when n is large, ML estimators $\hat{\theta}$ are *unbiased* with variance $\frac{I(\theta)^{-1}}{n}$. Hence ML estimators are efficient when n is large.

Relative efficiency:

- $Eff(\tilde{\theta}, \hat{\theta}) = \frac{Var(\hat{\theta})}{Var(\tilde{\theta})}$ — allows for comparison without knowing the exact Fisher information
- $Eff(\tilde{\theta}, \hat{\theta}) = Eff(\tilde{\theta})$ if $\hat{\theta}$ is efficient
- If relative efficiency > 1, then the first estimator $\tilde{\theta}$ is more efficient;

otherwise the second estimator $\hat{\theta}$ is more efficient

Efficiency of consistent estimators

- $Eff(\hat{\theta}) = \frac{I(\theta)^{-1}/n}{Var(\hat{\theta})}$
- $Eff(\tilde{\theta}, \hat{\theta}) = \frac{Var(\hat{\theta})}{Var(\tilde{\theta})}$
- ($\star$) Now it is possible for $Eff(\hat{\theta}) > 1$ for some values of n

11.3 Bias-Variance Tradeoff

- For an estimator to have lower variance than CRLB, it has to "pay" in terms of bias
- For an estimator to be unbiased, it has to "pay" in terms of variance

11.4 Choice of Estimator

How should we choose between estimators (e.g. MOM vs ML estimator)? What is the most important criteria?

- One approach: choose the one the lowest $MSE = E(\hat{\theta} - \theta)^2 = Var(\hat{\theta}) + (E(\hat{\theta}) - \theta)^2 = SE^2 + bias^2$
- Since $MSE = SE^2 + bias^2$, choosing smaller MSE means reducing SE and bias in some way
- But not always: sometimes, *unbiased* estimators are the most important

12 Sufficiency

12.1 Sufficiency and Factorisation Theorem

Sufficiency: T is *sufficient* for θ if the conditional distribution is the same across $\theta \in \Theta$, i.e. conditional distibution of $\boldsymbol{X}$ does not depend on θ for all possible values of $T = t$

T is *sufficient* for $\theta \leftrightarrow$ there is a function $q(\boldsymbol{x})$ such that for every $\theta \in \Theta$ and t,

$$f_\theta(\boldsymbol{X} = \boldsymbol{x} \mid T = t) = q(\boldsymbol{x}),\ \boldsymbol{x} \in S_t$$

(i.e. conditional distribution doesn't depend on θ, only $\boldsymbol{x}$)

Factorisation Theorem: T is *sufficient* for $\theta \leftrightarrow$ there exist functions $g(t, \theta)$ and $h(\boldsymbol{x})$ such that for every $\theta \in \Theta$ and t, $f_\theta(\boldsymbol{x}) = g(T(\boldsymbol{x}), \theta) \cdot h(\boldsymbol{x})$ for all possible $\boldsymbol{x}$

($\star$) Again, $f_\theta(\boldsymbol{x}) = g(T(\boldsymbol{x}), \theta) \cdot h(\boldsymbol{x})$

Examples of Factorisation:

- Bernoulli: $f(\boldsymbol{x}) = f(x_1, \ldots, x_n) = p^{\sum_i x_i}(1-p)^{n-\sum_i x_i}$, so $T = \sum_i X_i$ is sufficient
- Poisson: $f(\boldsymbol{x}) = f(x_1, \ldots, x_n) = (e^{-n\lambda}\lambda^{\sum_i x_i}) \cdot \frac{1}{\prod_i x_i!}$, so $T = \sum_i X_i$ is sufficient
- Normal: $f(\boldsymbol{x}) = f(x_1, \ldots, x_n) = \left(\frac{1}{\sqrt{2n}\sigma}\right)^n e^{\frac{\mu}{\sigma}\sum_i x_i - \frac{n\mu^2}{2\sigma^2}} e^{-\frac{1}{2\sigma^2}\sum_i x_i^2}$, so $T = (\sum_i X_i, \sum_i X_i^2)$ is sufficient

Intuition: T is *sufficient* if it can be used to *summarize* our samples $\boldsymbol{X}$, in a way that does not lose any information in our estimation of θ. E.g. Summing n IID Bernoullis loses no information in estimating parameter p

12.2 Conditional Expectations

Important facts

1. $E(Y) = E[E(Y|X)]$
2. $Var(Y) = Var[E(Y|X)] + E[Var(Y|X)]$

12.3 Rao-Blackwell Theorem

Take old estimator $\hat{\theta}$, and some sufficient statistic T. Define $\tilde{\theta} = E(\hat{\theta}|T)$. Then $\tilde{\theta}$ is a superior estimator with smaller MSE: $E(\tilde{\theta} - \theta)^2 \leq E(\hat{\theta} - \theta)^2$

13 Hypothesis Testing

Size and power

- Size: P(reject H_0 under H_0) — want this to be small
- Power: P(reject H_0 under H_1) — want this to be large

Neyman-Pearson approach: control size, maximise power

- Among tests with size $\leq \alpha$, choose the *most powerful* test

Likelihood ratio test

- Likelihood ratio: $\Lambda(\boldsymbol{x}) = \frac{\prod_i f_0(x_i)}{\prod_i f_1(x_i)} = \frac{f_0(x_1)\cdot\ldots\cdot f_0(x_n)}{f_1(x_1)\cdot\ldots\cdot f_1(x_n)}$
- Generalised likelihood ratio: $\Lambda = \frac{\max_{\theta\in\omega_0} L(\theta)}{\max_{\theta\in\Omega} L(\theta)}$ where $\omega_0 = \{\mu \mid \mu \text{ as in } H_0\}$ and $\Omega = \omega_0 \cup \omega_1$
- Let critical regions be of form $\{\boldsymbol{x} \mid \Lambda(\boldsymbol{x}) < c\}$, where $c > 0$

Neyman-Pearson lemma

- Most powerful test among tests with size $\leq \alpha$: has critical region $\{\boldsymbol{x} \mid \Lambda(\boldsymbol{x}) < c_\alpha\}$
- i.e. likelihood ratio test is the most powerful test
- Neyman-Pearson lemma gives a recipe for the *most powerful* test of size $\leq \alpha$ in the case of *simple* null and alternative hypotheses,

and the *uniformly most powerful* test in some cases of *composite* alternative hypotheses, but not in general

13.1 Pearson's Chi-Squared Statistic

$X^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$

- O_i: number of observations of type i
- E_i: expected number of observations of type i

Pearson's Chi-Squared Test:

- Rejection region: $X^2 > \chi^2_{k,\alpha}$ where k is the number of degrees of freedom

14 Comparing Two Samples: Independent Samples

Setup: X and Y values are independent, i.e. X_i and Y_i separately

- Assume independence, i.e. all X_i and Y_i are independent

Test $H_0 : \mu_X - \mu_Y = d$

14.1 Normal Theory: Same Variance σ^2

If σ^2 is known

Perform Z test

- Test statistic: $Z = \frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}}$
- SE of test statistic: $\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}$
- Reject H_0 when $|Z| > z_{\alpha/2}$ (two-tailed)
- $(1-\alpha)$ CI: $(\bar{X} - \bar{Y}) \pm z_{\alpha/2} \cdot \sigma\sqrt{\frac{1}{n} + \frac{1}{m}}$ (two-tailed)

If σ^2 is unknown

Estimate σ^2 using *pooled sample variance* s_p^2, an unbiased estimator

- $s_x^2 = \frac{1}{n-1}\sum_i (X_i - \bar{X})^2$
- $s_y^2 = \frac{1}{m-1}\sum_i (Y_i - \bar{Y})^2$
- $s_p^2 = \frac{(n-1)s_x^2 + (m-1)s_y^2}{m+n-2}$ — follows χ^2_{m+n-2} distribution

Perform T test

- t-statistic: $t = \frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{s_p\sqrt{\frac{1}{n} + \frac{1}{m}}}$ — t distribution with $m + n - 2$ DOF
- SE of t-statistic: $\sigma\sqrt{\frac{1}{n} + \frac{1}{m}} \approx s_p\sqrt{\frac{1}{n} + \frac{1}{m}}$

- Reject H_0 when t-statistic $> t_{m+n-2,\alpha/2}$ (two-tailed)
- $(1-\alpha)$ CI: $(\bar{X}-\bar{Y}) \pm t_{m+n-2,\alpha/2} \cdot s_p\sqrt{\frac{1}{n}+\frac{1}{m}}$ (two-tailed)
- (Note: with large n and m, approximately normal so just perform Z test)

14.2 Normal Theory: Different Variance σ_X^2 and σ_Y^2

If σ_X^2 and σ_Y^2 are known

$Var(\bar{X}-\bar{Y}) = \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}$

$Z = \frac{\bar{X}-\bar{Y}-(\mu_X-\mu_Y)}{\sqrt{\frac{\sigma_X^2}{n}+\frac{\sigma_Y^2}{n}}}$, still standard normal

If σ_X^2 and σ_Y^2 are unknown

Estimate them with s_X^2 and s_Y^2

$t = \frac{\bar{X}-\bar{Y}-(\mu_X-\mu_Y)}{\sqrt{\frac{s^2}{n}+\frac{s^2}{n}}}$, approximately t-distributed with DOF df

- $df = \frac{(a+b)^2}{\frac{a^2}{n-1}+\frac{b^2}{m-1}}$, where $a = \frac{s_X^2}{n}$, $b = \frac{s_Y^2}{n}$

14.3 Non-Parametric Test: Mann-Whitney (Wilcoxon) Test

Idea: under H_0, the ranks should be uniformly distributed, so rank sums should not be too small or too large.

- Data: Let $(Z_1, \ldots, Z_{m+n})$ be the pooled sample of X and Y values, and assume the values are distinct.
  - If there are tied values, assign them the *average* of the ranks
- Rank: Let $Rank(Z) = i$ if Z has i-th *smallest* value within the pooled sample
- Rank sum: let $T_X = \sum_{i=1}^{n} Rank(X_i)$, let $T_Y = \sum_{i=1}^{m} Rank(Y_i)$
  - Note that $T_X + T_Y = \sum_{i=1}^{m+n} i = \frac{(m+n)(m+n+1)}{2}$ is fixed
  - Take the smaller sample of size $n_1 = min(m,n)$, and compute sum of ranks R from that.
- Let $R' = n_1(m+n+1) - R$
- Let $R^* = min(R, R')$
- Reject $H_0 : F = G$ if R^* is too small

15 Comparing Two Samples: Paired Samples

Setup: X and Y values are paired up, i.e. (X_i, Y_i)

- Assume independence, i.e. (X_i, Y_i) and (X_j, Y_j) are independent
- Let $D_i = Y_i - X_i$

Test $H_0 : \mu_D = d$

15.1 Normal Theory

Perform T test

- t-statistic $t = \frac{\bar{D}-d}{s_D/\sqrt{n}}$, where $\bar{D}$ is sample mean and s_D^2 is sample variance

- Reject H_0 when $|t| > t_{n-1,\alpha/2}$ (two-tailed test)
- $(1-\alpha)$ CI: $\bar{D} \pm t_{n-1,\alpha/2} \cdot s_D/\sqrt{n}$
- (Note: with large n, approximately normal so just perform Z test)

15.2 Non-Parametric Test: Wilcoxon Signed-Rank Test

Idea: under H_0, distribution of D_i is symmetrically distributed around 0

- Data: Let $D_1, \ldots, D_n$ be the sample of differences
- Rank: Let $Rank(D) = i$ if D has i-th smallest absolute value in the sample
- Rank sum: Let W_+ be rank sum among positive D_i, let W_- be rank sum among negative D_i
  - Note that $W_+ + W_- = 1 + \ldots + n = \frac{n(n+1)}{2}$ is fixed
- Let $W = min(W_+, W_-)$
- Reject H_0 if W is too small