Lecture Notes

[2020-08-12 Wed]

Contents

1 Assessment

Final Exam	70%
Assignments (x4)	30%

2 Introduction and Recap

2.1 Independence

Independence: X_1 and X_2 are independent if for any 2 events A_1 and A_2,

- $P(X_1 \in A_1, X_2 \in A_2) = P(X_1 \in A_1) \cdot P(X_2 \in A_2)$
- $P(A_1 \cap A_2) = P(A_1) \cdot P(A_2)$

2.2 Expectation and Variance

Expectation: $E(X) = \int x \cdot f_X(x)\ dx$

- $E(g(X)) = \int g(x) \cdot f_X(x)\ dx$
- $E(aX + bY) = aE(X) + bE(Y)$

Variance: $Var(X) = E[X - E(X)^2] = E(X^2) - E(X)^2$

- $Var(aX + bY) = a^2Var(X) + b^2Var(Y) + 2ab\ Cov(X, Y)$

Standard deviation: $SD(X) = \sqrt{Var(X)}$

Covariance: $Cov(X, Y) = E[(X - E(X))(Y - E(Y))] = E(XY) - E(X)E(Y)$

- $Cov(aX, bY) = ab\ Cov(X, Y)$
- $Cov(X + a, Y + b) = Cov(X, Y)$

Correlation: $p_{X,Y} = \frac{Cov(X,Y)}{SD(X)SD(Y)}$

2.3 Centred Random Variables

Centred: zero expectation

We can centre RV X by letting $Y = X - E(X)$

- $E(Y) = 0$
- $Var(Y) = Var(X)$

Model of measurement: Let μ be a constant (what we're interested in) and ϵ be centred RV (random measurement errors), define $X = \mu + \epsilon$ (measurements)

- $E(X) = \mu$
- $Var(X) = Var(\epsilon)$
- $Cov(X, \epsilon) = Cov(c + \epsilon, \epsilon) = Cov(\epsilon, \epsilon) = Var(\epsilon)$

2.4 Law of Large Numbers

Law of large numbers: Let X_i be IIDs with expectation μ and variance σ^2, and mean $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Then as $n \to \infty$, $\bar{X}_n \to \mu$.

- Implication: Let x_i be realisations of X_i, and so $\bar{x}_n$ is a realisation of $\bar{X}_n$. Then as $n \to \infty$, $\bar{x}_n \to \mu$; variance of x_i converges to σ^2.
- Empirical distribution converges to the common distribution

2.5 Central Limit Theorem

Let $X_1, \ldots, X_n$ be IID (any distribution!) with expectation μ and SD σ. Let $S_n = \sum_i X_i$.

($\star$) As $n \to \infty$, distribution of $\frac{S_n - n\mu}{\sqrt{n}\sigma}$ converges to standard normal distribution, then $S_n \sim N(n\mu, n\sigma^2)$

Example: 100 tosses of fair coin, find P(between 40 and 60 heads)

- $S_n \sim Bin(100, \frac{1}{2}) \sim> N(50, 25)$
- $P(40 \leq S_n \leq 60) \approx P(\frac{40-50}{5} < Z < \frac{60-50}{5}) \approx 95\%$ (ignoring continuity correction)

3 Normal Distribution and Related Distributions

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

- $f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
- $E(X) = \mu$
- If $\mu = 0$, then $E(X^{2i+1}) = 0$

3.1 Standardisation

Let $Z = \frac{X-\mu}{\sigma}$, then Z is standard normal

- Z has pdf ϕ and cdf Φ

3.2 χ^2 Distribution

χ_1^2 Distribution

Let $U = Z^2$. Then $U \sim \chi_1^2$

- $F_U(u) = \Phi(\sqrt{u}) - \Phi(-\sqrt{u})$
- $f_U(u) = \frac{1}{\sqrt{2\pi}} u^{-\frac{1}{2}} e^{-\frac{u}{2}}$
- $E(U) = 1$
- $Var(U) = 2$

χ_1^2 is Gamma with $\alpha = \frac{1}{2}, \lambda = \frac{1}{2}$

(Proof) Find CDF of U, $F_U(u)$

$$\begin{aligned} F_U(u) = P(U \le u) &= P(Z^2 \le u) \\ &= P(-\sqrt{u} \le Z \le \sqrt{u}) \\ &= P(Z \le \sqrt{u}) - P(Z \le -\sqrt{u}) \\ &= \Phi(\sqrt{u}) - \Phi(-\sqrt{u}) \end{aligned}$$

$$\begin{aligned} f_U(u) &= \frac{d}{du} F_U(u) \\ &= \phi(\sqrt{u}) \frac{1}{2} u^{-\frac{1}{2}} - \phi(-\sqrt{u})(-\frac{1}{2} u^{-\frac{1}{2}}) \\ &= \phi(\sqrt{u}) \frac{1}{2} u^{-\frac{1}{2}} + \phi(\sqrt{u})(\frac{1}{2} u^{-\frac{1}{2}}) \\ &= \phi(\sqrt{u}) u^{-\frac{1}{2}} \\ &= \frac{1}{\sqrt{2\pi}} u^{-\frac{1}{2}} e^{-\frac{u}{2}} \end{aligned}$$

χ_n^2 Distribution

Let $V = \sum_{i=1}^n U_i$, where $U_1, \ldots, U_n$ are IID χ_1^2. Then $V \sim \chi_n^2$

- χ_n^2 is Gamma with $\alpha = \frac{n}{2}$, $\lambda = \frac{1}{2}$

- $E(V) = n$
- $Var(V) = 2n$
- 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 $X + Y \sim \chi^2_{m+n}$

3.3 Γ Distribution

$g(t) = \frac{\lambda^\alpha}{\Gamma(\alpha)} t^{\alpha-1} e^{-\lambda t}$, for $t \geq 0$

- α is the *shape* parameter
- λ is the *rate* parameter — how fast it 'dies out'
- If we have independent Gamma distributions with the same rate parameter, we can simply sum the shape paremeters!

3.4 Moment Generating Functions

$M_V(t) = E(e^{tV})$

- From $M(t)$, you can get all the moments of V, e.g. $E(V)$, $E(V^2)$!
- $E(V^k) = M^{(k)}(0)$, so $E(V) = M'(0)$, $E(V^2) = M''(0)$

($\star$) MGF uniquely describes the distribution. If two variables have the same MGF, they have the same distribution.

($\star$) If X and Y are independent, then $M_{X+Y}(t) = M_X(t) \cdot M_Y(t)$

MGF of Γ and χ^2_n distributions

- For Γ distribution, $M(t) = (\frac{\lambda}{\lambda - t})^\alpha$
- For χ^2_n distribution, $M(t) = (1-2t)^{-\frac{n}{2}}$

Expectation and Variance of χ^2_n distribution

$$
\begin{aligned}
M(t) &= (1-2t)^{-\frac{n}{2}} \\
M'(t) &= n(1-2t)^{-\frac{n}{2}-1} \\
M''(t) &= n(n+2)(1-2t)^{-\frac{n}{2}-2}
\end{aligned}
$$

$$
\begin{aligned}
E(V) &= M'(0) = n \\
E(V^2) &= M''(0) = n(n+2) \\
Var(V) &= E(V^2) - E(V)^2 = 2n
\end{aligned}
$$

Proving that Sum of Chi-Squares gives Chi-Square with $n+m$ Degrees of Freedom

Since X and Y are independent,

$$
\begin{aligned}
M_{X+Y}(t) &= M_X(t)M_Y(t) \text{ (since } X \text{ and } Y \text{ are independent)} \\
&= (1-2t)^{-\frac{m}{2}}(1-2t)^{-\frac{n}{2}} \\
&= (1-2t)^{-\frac{m+n}{2}}
\end{aligned}
$$

$$\therefore X + Y \sim \chi^2_{m+n}$$

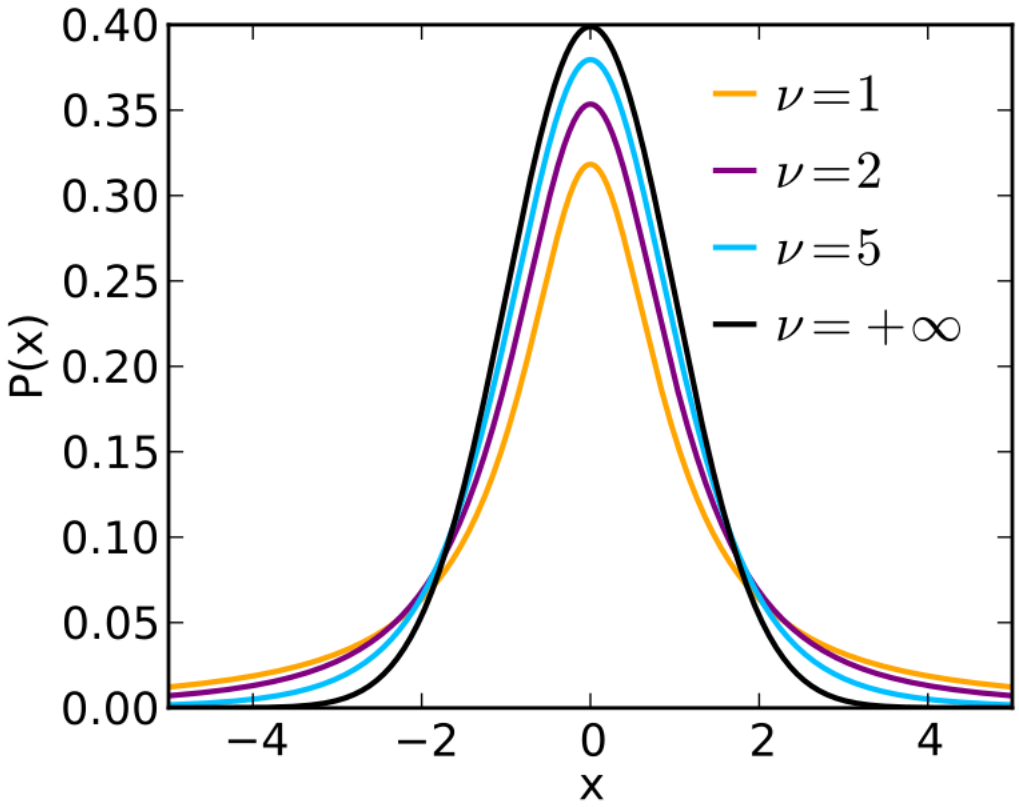
3.5 t Distribution

$t_n = \frac{Z}{\sqrt{V_n/n}}$, where $V_n \sim \chi^2_n$

- t_n has a t distribution with n degrees of freedom
- $f(t) = \frac{\Gamma((n+1)/2)}{\sqrt{n\pi}\Gamma(n/2)}(1+\frac{t^2}{n})^{-(n+1)/2}$

t_n distribution

- Symmetric about 0
- Converges to Z as $n \to \infty$
  - V_n/n: by law of large numbers, as $n \to \infty$, V_n/n converges to its mean 1, so t_n converges to Z



Deriving PDF of t Distribution

(Not important)

$$X = \sqrt{\frac{V_n}{n}}$$

$$\begin{aligned}
P(X \le x) &= P(\sqrt{\frac{V_n}{n}} \le x) \\
&= P(V_n \le nx^2) \\
f_X(x) &= f_{V_n}(nx^2) \cdot (2nx) \\
&= \frac{(\frac{1}{2})^{\frac{n}{2}}}{\Gamma(\frac{n}{2})}(nx^2)^{\frac{n}{2}-1}e^{-\frac{1}{2}nx^2}(2nx) \\
&= \frac{1}{\Gamma(\frac{n}{2})}(\frac{1}{2})^{\frac{n}{2}-1}n^{\frac{n}{2}}x^{n-1}e^{-\frac{1}{2}nx^2} \\
&= Kx^{n-1}e^{-\frac{1}{2}nx^2}
\end{aligned}$$

$$\begin{aligned}
\text{Theorem: } f_Z(z) &= \int_{-\infty}^{\infty} |x| f_X(x) f_Y(xz)\ dx \text{ when } Z = \frac{Y}{X} \\
\text{So: } f_t(t) &= \int_{-\infty}^{\infty} |x| f_X(x) \psi(xt)\ dx \text{ since } t = \frac{Z}{X} \\
&= \int_0^{\infty} xKx^{n-1}e^{-\frac{1}{2}nx^2}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2t^2}\ dx \\
&= \frac{K}{\sqrt{2\pi}}\int_0^{\infty} x^n e^{-\frac{1}{2}x^2(n+t^2)}\ dx \\
&= \frac{K}{\sqrt{2\pi}}\int_0^{\infty} \frac{1}{2}y^{\frac{n-1}{2}}e^{-\frac{1}{2}y(n+t^2)}\ dy \text{ (Let } y = x^2, \frac{dy}{dx} = 2x) \\
&\text{Consider Gamma density function with } \alpha = \frac{n+1}{2}, \lambda = \frac{n+t^2}{2} \\
&= \frac{K}{\sqrt{2\pi}} \cdot \frac{1}{2} \cdot \frac{\Gamma(\alpha)}{\lambda^\alpha}\int_0^{\infty} \frac{\lambda^\alpha}{\Gamma(\alpha)}y^{\alpha-1}e^{-\lambda y}\ dy \text{ (The integral is 1, because Gamma!)} \\
&= \ldots = k \cdot (1 + \frac{t^2}{n})^{-(n+1)/2}
\end{aligned}$$

3.6 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
- Note: 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)
- If $n \to \infty$, then $W \approx U/m$ since $\frac{V}{n} \to 1$
- For $n > 2$, $E(W) = \frac{n}{n-2}$

3.7 ($\star$) IID Normal Random Variables: Sample Mean $\bar{X}$ and "Sample Variance" S^2

Let $X_1, \ldots, X_n$ be IID $N(\mu, \sigma^2)$.

- Sample mean: $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$

- "Sample variance": $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ — NOTE the $(n-1)$!

Facts about $\bar{X}$ and S^2:

1. $\bar{X}$ and S^2 are independent
2. $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$
3. $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$
4. $\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}$
5. $E(S^2) = \sigma^2$ — since $E(\frac{(n-1)S^2}{\sigma^2}) = n-1$
6. $Var(S^2) = \frac{2\sigma^4}{n-1}$ — since $Var(\frac{(n-1)S^2}{\sigma^2}) = 2(n-1)$

Proof of Fact 1 ($\bar{X}$ and S^2 are Independent)

- Theorem: If X and Y are independent, then $E(XY) = E(X)E(Y)$; if (X, Y) is bivariate/multivariate normal, then the converse is also true
- Idea: show that $E(\bar{X}Y) = E(\bar{X})E(Y)$, where $Y = X_1 - \bar{X}$; hence $\bar{X}$ and $X_1 - \bar{X}$ are independent (by theorem), and hence $\bar{X}$ and S^2 are independent

$$\begin{aligned}
E(\bar{X}Y) = E(\bar{X}(X_1 - \bar{X})) &= E(X_1(\frac{1}{n}(X_1 + \ldots + X_n))) - E(\bar{X}^2) \\
&= \frac{1}{n}E(X_1^2 + X_1X_2 + \ldots + X_1X_n) - E(\bar{X}^2) \\
&= \frac{1}{n}[E(X_1^2) + E(X_1)E(X_2) + \ldots + E(X_1)E(X_n)] - E(\bar{X}^2) \\
&= \frac{1}{n}[(\sigma^2 + \mu^2) + (n-1)\mu^2] - E(\bar{X}^2) \\
&= \frac{\sigma^2}{n} + \mu^2 - E(\bar{X}^2) \\
&= \frac{\sigma^2}{n} + \mu^2 - [Var(\bar{X}) + E(\bar{X})^2] \\
&= \frac{\sigma^2}{n} + \mu^2 - [\frac{\sigma^2}{n} + \mu^2] \\
&= 0
\end{aligned}$$

$$E(\bar{X})E(Y) = \mu \cdot E(X_1 - \bar{X}) = \mu \cdot 0 = 0$$

Proof of Fact 2 (Distribution of $\bar{X}$)

Obvious: linear combination of IID normal RVs is a normal RV

Proof of Fact 3 (Distribution of S^2)

Note that:

$\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2 = \sum_{i=1}^{n}(\frac{X_i-\mu}{\sigma})^2 \sim \chi^2_n$

Verify:

$\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2 = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \bar{X})^2 + (\frac{\bar{X}-\mu}{\sigma/\sqrt{n}})^2$

$$\sum_i (X_i - \mu)^2 = \sum_i ((X_i - \bar{X}) + (\bar{X} - \mu))^2$$
$$= \sum_i (X_i - \bar{X})^2 + \sum_i (\bar{X} - \mu)^2$$

Call these terms $U = V + W$.

U and V are independent from Fact 1, so:

$M_W(t) = \frac{M_W(t)}{M_V(t)} = \frac{(1-2t)^{-n/2}}{(1-2t)^{-n/2}} = (1-2t)^{-(n-1)/2}$

and U has a χ^2_{n-1} distribution

Proof of Fact 4 (t Distribution)

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} = \frac{(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}})}{(\frac{S/\sqrt{n}}{\sigma/\sqrt{n}})}$$

Numerator: $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$

Denominator: $\frac{S}{\sigma} = \sqrt{\frac{S^2}{\sigma^2} \cdot \frac{(n-1)}{(n-1)}} = \sqrt{\frac{S^2(n-1)}{\sigma^2}/(n-1)}$

- It has χ^2_{n-1} distribution divided by $(n-1)$, square rooted.

Conclusion: $\frac{\bar{X}-\mu}{S/\sqrt{n}}$ has same distribution as $\frac{Z}{\sqrt{U_{n-1}/(n-1)}}$, which is t_{n-1} distribution

3.8 ($\star$) IID General Random Variables: Sample Mean $\bar{X}$ and "Sample Variance" S^2

Now *relax* the normality assumption. Let $X_1, \ldots, X_n$ be IID with expectation μ and variance σ^2.

As $n \to \infty$, $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$ (by CLT), so for large n, $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$

4 Survey Sampling A: SRS, Estimation

4.1 Motivation

Population is too large to study; so we choose a smaller sample, and use it to infer facts about the population

- *Random* samples: uses chance to gather sample, the best way to choose sample

4.2 Population

Let population have size N, each member has fixed value x_i. Let μ, τ, σ be mean, total, SD of the population respectively.

- $\mu = \frac{1}{N}\sum_i x_i$
- $\tau = \sum_i x_i$
- $\sigma^2 = \frac{1}{N}\sum_i (x_i - \mu)^2 = \frac{1}{N}\sum_i x_i^2 - \mu^2$ (the $\frac{1}{N}$ is *out*)
- $\sigma = \sqrt{\frac{1}{N}\sum_i (x_i - \mu)^2}$

($\star$) The mean and variance are now no longer defined based on *random variables*, but on *a list of numbers*! So they are not random at all, just a bunch of numbers.

4.3 Population of 0's and 1's

Each member can take value only 0 or 1, let p be the proportion of the population that take the value 1

- $\mu = p$
- $\sigma^2 = p(1-p)$
- $\sigma = \sqrt{p(1-p)}$

4.4 Simple Random Sampling

Simple random sampling: make n random draws *without replacement* from the population

- Every subset of size n has the same probability

Represent the draws as $X_1, \ldots, X_n$

- X_i's have the same (marginal) distribution
- $E(X_i) = \mu$, $Var(X_i) = \sigma^2$
- X_i and X_j are dependent and negatively correlated, $Cov(X_i, X_j) = -\frac{\sigma^2}{N-1}$

Proof of $E(X_i)$ and $Var(X_i)$ (Lemma A)

Let the *values* of the population members be $\xi_1, \ldots, \xi_m$, let the *number* of each value ξ_j be n_j.

$P(X_i = \xi_j) = \frac{n_j}{N}$

$E(X_i) = \sum_{j=1}^m \xi_j P(X_i = \xi_j) = \sum_{j=1}^m \xi_j \frac{n_j}{N} = \frac{1}{N}\sum_{j=1}^m n_j \xi_j = \mu$

$Var(X_i) = \sum_{j=1}^m (\xi_j - \mu)^2 P(X_i = \xi_j) = \frac{1}{N}\sum_{j=1}^m n_j (\xi_j - \mu)^2 = \frac{1}{N}\sum_{i=1}^N (x_i - \mu)^2 = \sigma^2$

Proof of $Cov(X_i, X_j)$ (Lemma B)

$Cov(X_i, X_j) = E(X_i X_j) - E(X_i)E(X_j)$

$$\begin{aligned}
E(X_i X_j) &= \sum_{k=1}^{m}\sum_{l=1}^{m} \xi_k \xi_l P(X_i = \xi_k, X_j = \xi_l) \\
&= \sum_{k=1}^{m} \xi_k P(X_i = k) \sum_{l=1}^{m} \xi_l P(X_j = \xi_l \mid X_i = \xi_k) \\
&= \sum_{k=1}^{m} \xi_k \frac{n_k}{N} \left[\sum_{l \neq k} \xi_l \frac{n_l}{N-1} + \xi_k \frac{n_l - 1}{N-1} \right] \\
&= \sum_{k=1}^{m} \xi_k \frac{n_k}{N} \left[\sum_{l=1}^{m} \xi_l \frac{n_l}{N-1} - \xi_k \frac{1}{N-1} \right] \\
&= \frac{1}{N(N-1)} \tau(\tau - \xi_k) \\
&= \frac{\tau}{N(N-1)} - \frac{1}{N-1} \left(\frac{1}{N} \sum_{k=1}^{m} \xi_k^2 n_k \right) \\
&= \frac{\tau}{N(N-1)} - \frac{1}{N-1}(\mu^2 + \sigma^2) \\
&= \dots \\
&= \mu^2 - \frac{\sigma^2}{N-1}
\end{aligned}$$

$$P(X_j = \xi_l \mid X_i = \xi_k) = \begin{cases} \frac{n_l}{N-1} & \text{if } k \neq l \\ \frac{n_l - 1}{N-1} & \text{if } k = l \end{cases}$$

$$\frac{1}{N} \sum_{k=1}^{m} \xi_k n_k = E(X_i) = \mu$$
$$\frac{1}{N} \sum_{k=1}^{m} \xi_k^2 n_k = E(X_i^2) = Var(X_i) + E(X_i)^2 = \mu^2 + \sigma^2$$

4.5 Simple Random Sampling Facts

Let $\bar{X}$ be the sample mean of $X_1, \dots, X_n$, and let $T = N\bar{X}$

1. $E(\bar{X}) = \mu$, $E(T) = \tau$
2. $Var(\bar{X}) = \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1}$
3. $Var(T) = N^2 \cdot \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1}$

Proof of Fact 2

$$\begin{aligned} Var(\bar{X}) &= \frac{1}{n^2}\left[\sum_{i=1}^{n} Var(X_i) + \sum_{i\neq j} Cov(X_i, X_j)\right] \\ &= \frac{1}{n^2}\left[n\sigma^2 + n(n-1)\cdot\frac{-\sigma^2}{N-1}\right] \\ &= \frac{\sigma^2}{n}\left[1 - \frac{n-1}{N-1}\right] \\ &= \frac{\sigma^2}{n}\cdot\frac{N-n}{N-1} \end{aligned}$$

4.6 Finite Population Correction Factor

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

- It's smaller than 1
- So SD is smaller than for draws with replacement

Sampling fraction: $\frac{n}{N}$

- When sampling fraction is small, the finite population correction is close to 1

4.7 Special case: Proportion (0/1)

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

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

5 Estimation Problem

Suppose that population mean μ is unknown. Let $X_1, \ldots, X_n$ be random draws *with replacement*.

- $\bar{X}$ is an *estimator* of μ
- $\bar{x}$, an observed value of $\bar{X}$, an *estimate* of μ

Conclusion

- $\bar{x}$ is an unbiased estimate of μ
- s^2 is an unbiased estimate of σ^2

5.1 Standard Error (SE)

SE (of $\bar{x}$) $= SD(\bar{X}) = \frac{\sigma}{\sqrt{n}}$

- Error in a particular estimate $\bar{x}$ is unknown, but on average its size is SE
- SE is independent of population size N, dependent on sample size n
- SE is unknown since we don't know σ or σ^2

5.2 $\hat{\sigma}^2$ (the Bootstrap) as a biased estimator of σ^2

Let's try $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$:

- Since $\sigma^2 = E((X_i - \mu)^2) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$ — simply replace μ by $\bar{X}$
- However, $\hat{\sigma}^2$ is a biased estimator: $E(\hat{\sigma}^2) = (\frac{n}{n-1})\sigma^2$

$$
\begin{aligned}
E(\sum_i (X_i - \bar{X})^2) &= E(\sum_i X_i + n\bar{X}^2 - 2\bar{X}\sum_i X_i) \\
&= E(\sum_i X_i^2 - n\bar{X}^2) \\
&= \sum_i E(X_i^2) - nE(\bar{X}^2) \\
&= \sum_i \left(Var(X_i) + E(X_i)^2\right) - n(Var(\bar{X}) + E(\bar{X})^2) \\
&= n\left(\sigma^2 + \mu^2\right) - n(\frac{\sigma^2}{n} + \mu^2) \\
&= \frac{n}{n-1}\sigma^2
\end{aligned}
$$

5.3 S^2 as an unbiased estimator of σ^2, and s^2 as an unbiased estimate

$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{X})^2$

- $E(S^2) = \sigma^2$

$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$, where s^2 is an estimate of S^2

5.4 Special Case: Proportion of 0/1

Let's say we have $\hat{p}$. Then $\hat{\sigma}^2 = \hat{p}(1-\hat{p})$, $s^2 = \frac{n}{n-1}\hat{\sigma}^2$

So approximate SE of $\hat{p}$ is $\sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}}$

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

5.5 Simple Random Sampling (Without Replacement)

Without replacement, we have to re-introduce the finite population correction factor: multiply the SE by $\sqrt{1-\frac{n-1}{N-1}}$

S^2 is biased for σ^2, so $E\left(\frac{N-1}{N}S^2\right) = \sigma^2$

But practically, we don't do this correction: the correction factor has a negligible effect when N is large compared to n

6 Survey Sampling B: Confidence Intervals, Measurement Model

6.1 Normal Approximation for $\bar{X}$

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

6.2 Confidence Intervals

Let $x_1, \ldots, x_n$. Let z_u be such that $P(Z > z_u) = u$.

- E.g. $z_{0.025} = 1.96$, because $P(Z > 1.96) = 0.025$

Let α be the significance level. Let the \$(1 - α)\$-CI for μ be:

$(\bar{x} - z_{\alpha/2}\frac{s}{\sqrt{n}}, \bar{x} + z_{\alpha/2}\frac{s}{\sqrt{n}})$

which is a realisation of the random interval:

$(\bar{X} - z_{\alpha/2}\frac{s}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{s}{\sqrt{n}})$

whereby $1 - \alpha \approx P(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}})$.

Hence the probability of a CI containing μ is $(1 - \alpha)$.

Example

Say we take 100 samples from a population of 8000. $\bar{x} = 1.6$, $s = 0.8$. Construct a 95%-CI for μ.

- Approximate SE is $\frac{s}{\sqrt{n}} = \frac{0.8}{\sqrt{100}} = 0.08$
- $\alpha = 0.05$, and $z_{\alpha/2} = 1.96$
- So approximate 95%-CI for μ is $1.6 \pm 1.96 \cdot 0.08 \approx (1.4, 1.8)$

Example

Say 12% plans to sell house unit next year. Construct a 95%-CI for p.

- Approximate SE is $\sqrt{\frac{0.12 \times 0.88}{99}} = 0.03$
- So approximate 95%-CI for p is $0.12 \pm 1.96 \times 0.03 = (0.06, 0.18)$

6.3 Summary on Estimation of Population Mean μ

- Simple Random Sampling (SRS): we can take it to be sampling with replacement (when n is much smaller than N)
- SE of estimate $\bar{x}$ is $SD(\bar{X}) = \sigma/\sqrt{n}$
- If sample size n is large, S^2 is similar to $\hat{\sigma}^2$
- If sample size n is large, then distribution of $\bar{X}$ is approximately normal; then interval (estimate $\pm$ 1.96 SE) is a good 95%-CI for μ

Exercise

Say we have 2 populations, population I and population II

- Population I: sample size n_1, population SD σ_1
- Population II: sample size $n_2 = 2n_1$, population SD $\sigma_2 = 2\sigma_1$

Which population has a larger confidence interval?

- Width of CI $W = 2z_{\alpha/2}\frac{s}{\sqrt{n}}$
- $\frac{W_1}{W_2} = \frac{S_1/\sqrt{n_1}}{S_2/\sqrt{n_2}} \approx \frac{\sigma_1/\sqrt{n_1}}{\sigma_2/\sqrt{n_2}} = \frac{\sigma_1/\sqrt{n_1}}{2\sigma_1/\sqrt{2n_1}} = \frac{1}{\sqrt{2}}$

Exercise

Say we have population $p = 0.2$, take sample size $n = 100$

1. Find δ such that $P(|\hat{p} - p| \geq \delta) = 0.025$.
2. If $\hat{p} = 0.25$, will the 95%-CI for p contain the true value of p?
3. $\hat{p} \sim N(p, \frac{p(1-p)}{n}) = N(0.2, \frac{1}{25^2})$
4. $P(|\hat{p} - 0.2| \geq \delta) = 0.025$
5. $P(|\frac{\hat{p}-0.2}{\frac{1}{25}}| \geq \frac{\delta}{\frac{1}{25}}) = 0.025$
6. $P(|Z| \geq 25\delta) = 0.025$
7. $25\delta = z_{0.0125} \Rightarrow \delta = \frac{1}{25}z_{0.0125}$

Then:

- $\hat{p} \pm z_{0.0125}\sqrt{\frac{\frac{1}{4}\frac{3}{4}}{99}} = 0.25 \pm 0.08$
- So 95%-CI contains true value of $p = 0.2$

6.4 Measurement Error

Often, we can assume that $X_i = \mu + \epsilon_i$, where errors ϵ_i are IID with expectation 0 and unknown variance σ^2.

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

Then $x_i = \mu + e_i$, where e_i is a realisation of ϵ_i

- Similarly, $\bar{x}$ is an estimate of μ, and its SE is $\frac{\sigma}{\sqrt{n}}$

6.5 Biased Measurements

Let $X = \mu + \epsilon$, and we want to use X to measure a constant $a \neq \mu$.

Then $X = a + (\mu - a) + \epsilon$, where the bias is $\mu - a$.

Mean square error $MSE = SE^2 + bias^2$

- The MSE for X, $MSE = E[(X - a)^2] = E[((\mu - a) + \epsilon)^2] = E[(\mu - a)^2 + \epsilon^2 + 2\epsilon(\mu - a)] = \sigma^2 + (\mu - a)^2$
- The MSE for $\bar{X}$, $MSE = E[(\bar{X} - a)^2] = \frac{\sigma^2}{n} + (\mu - a)^2$

Implication of biased measurements

- As you take more biased measurements, SE^2 will decrease, but $bias^2$ will remain the same
- If $a = \mu$, $\sqrt{MSE} = SE$, so SE indicates accuracy of $\bar{x}$
- If $a \neq \mu$, $\sqrt{MSE} > SE$, so SE only partly indicates accuracy of $\bar{x}$

7 Survey Sampling C: Estimation of a Ratio

Setup: Population of N members, record two characteristics from each member, $(x_1, y_1), \ldots, (x_N, y_N)$

- $\mu_X = \frac{1}{N}\sum_i x_i$
- $\mu_Y = \frac{1}{N}\sum_i y_i$

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

Estimator of ratio parameter can be $R = \frac{\bar{Y}}{\bar{X}}$

- Good estimator of r when n is large, because of law of large numbers for $\bar{X}$ and $\bar{Y}$
- $E(R)$?
- $Var(R)$? It should be small (see next section)

7.1 Approximating $Var(R)$ (via Taylor expansion)

- Let $f(x, y) = \frac{y}{x}$, so $r = f(\mu_x, \mu_y)$ and $R = f(\bar{X}, \bar{Y})$
- $\frac{\partial f}{\partial x}(x, y) = -\frac{y}{x^2}$
- $\frac{\partial f}{\partial y}(x, y) = \frac{1}{x}$

Assume that $\bar{X} \approx \mu_x$ and $\bar{Y} \approx \mu_y$ (when n is large):

Approximate up to first-order terms:

$$\begin{aligned} R &= f(\bar{X}, \bar{Y}) \\ &= f(\mu_x, \mu_y) + \frac{\partial f}{\partial x}(\mu_x, \mu_y)(\bar{X} - \mu_x) + \frac{\partial f}{\partial y}(\mu_x, \mu_y)(\bar{Y} - \mu_y) + \ldots \text{ (by Taylor expansion)} \\ &= r + (\frac{-\mu_y}{\mu_x^2})(\bar{X} - \mu_x) + (\frac{1}{\mu_x})(\bar{Y} - \mu_y) + \ldots \end{aligned}$$

$$\begin{aligned} Var(R) &= Var(...) \\ &\approx \frac{\mu_y^2}{\mu_x^4}\sigma_{\bar{X}}^2 + \frac{1}{\mu_x^2}\sigma_{\bar{Y}}^2 - 2\frac{\mu_y}{\mu_x^3}\sigma_{\bar{X}\bar{Y}} \\ &= \frac{1}{\mu_x^2}(r^2\sigma_{\bar{X}}^2 + \sigma_{\bar{Y}}^2 - 2r\sigma_{\bar{X}\bar{Y}}) \end{aligned}$$

7.2 Simple Random Sampling

Same finite population correction factor for SRS.

- $Var(\bar{X}) = \frac{\sigma_x^2}{n}(\frac{N-n}{N-1})$
- $Var(\bar{Y}) = \frac{\sigma_y^2}{n}(\frac{N-n}{N-1})$
- $Cov(\bar{X}, \bar{Y}) = \frac{\sigma_{xy}}{n}(\frac{N-n}{N-1})$
- Where $\sigma_{xy} = \frac{1}{N}\sum_i (x_i - \mu_x)(y_i - \mu_y)$

7.3 Approximating $E(R)$ (via Taylor expansion)

- $\frac{\partial f}{\partial x}(x,y) = -\frac{y}{x^2}$
- $\frac{\partial f}{\partial y}(x,y) = \frac{1}{x}$
- $\frac{\partial^2 f}{\partial x^2}(x,y) = \frac{2y}{x^3}$
- $\frac{\partial^2 f}{\partial y^2}(x,y) = 0$
- $\frac{\partial^2 f}{\partial x \partial y}(x,y) = -\frac{1}{x^2}$

Approximate up to second-order terms:

$$
\begin{aligned}
R &= f(\bar{X}, \bar{Y}) \\
&= f(\mu_x, \mu_y) + \frac{\partial f}{\partial x}(\mu_x, \mu_y)(\bar{X} - \mu_x) + \frac{\partial f}{\partial y}(\mu_x, \mu_y)(\bar{Y} - \mu_y) + \frac{1}{2}(\frac{2\mu_y}{\mu_x^3})(\bar{X} - \mu_x)^2 + \frac{1}{2}(0)(\bar{Y} - \mu_y)^2 + (-\frac{1}{\mu_x^2})(\bar{X} - \mu_x)(\bar{Y} - \\
&= r + (\frac{-\mu_y}{\mu_x^2})(\bar{X} - \mu_x) + (\frac{1}{\mu_x})(\bar{Y} - \mu_y) + \dots
\end{aligned}
$$

$$
\begin{aligned}
E(R) &= E(...) \\
&= E(r) + 0 + 0 + \frac{\mu_y}{\mu_x^3} E[(\bar{X} - \mu_x)^2] + 0 - \frac{1}{\mu_x^2} E[(\bar{X} - \mu_x)(\bar{Y} - \mu_y)] + \dots \\
&= r + \frac{1}{\mu_x}(r\sigma_{\bar{X}}^2 - \sigma_{\bar{X}\bar{Y}}) + \dots \\
&\approx r + \frac{1}{n}(\frac{N-n}{N-1})\frac{1}{\mu_x}(r\sigma_x^2 - \rho\sigma_x\sigma_y) \text{ (SRS)}
\end{aligned}
$$

7.4 Population Correlation Coefficient

$\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$

Unitless, takes values between -1 and 1

- Close to -1 and 1: strong linear relationship between x and y

7.5 Estimating σ

Estimate σ_x^2 using $s_x^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$

Estimate σ_{xy} using $s_{xy} = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})$

Estimate $\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \approx \frac{s_{xy}}{s_x s_y} = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_i (X_i - \bar{X})^2}\sqrt{\sum_i (Y_i - \bar{Y})^2}}$

7.6 Confidence Interval

How do we form a $(1-\alpha)$ confidence interval for r?

- Previously, we used $\bar{x} \pm z_{\frac{\alpha}{2}} SE(\bar{x})$ as approximate $(1-\alpha)$ CI for μ
- Now we try using $R \pm z_{\frac{\alpha}{2}} SE(R)$ as approximate $(1-\alpha)$ CI for r (here R refers to the *estimate*, not the estimator, a bit of an abuse of notation)
- This only works (using z) if R is approximately normal, which is true (by Taylor expansion)

– R is a linear combination of $\bar{X}$ and $\bar{Y}$, and they're roughly independent (unlike X and Y that might have covariance); so linear combination of jointly normal is approximately normal (technically multivariate CLT but don't need to know this)

- n must be large so that it's OK to ignore bias (the extra term in $E(R)$), and is approximately normal by CLT

We estimate R using the sample estimate for R, estimate $Var(R)$ using S_R^2 by plugging in sample estimates

- $S_R^2 = \frac{1}{n}(\frac{N-n}{N-1})\frac{1}{\bar{x}^2}(R^2 s_x^2 + s_y^2 - 2Rs_{xy})$ (with SRS finite population correction)
- Then $R \pm z_{\frac{\alpha}{2}} S_R$ is approx $(1-\alpha)$ CI for r.

7.7 Some quick summary

$\sigma_{\bar{X}}^2 = Var(\bar{X})$

$SE(\bar{x}) = SD(\bar{X}) = \sigma_{\bar{X}} = \frac{\sigma_x}{\sqrt{n}} = \frac{s_x}{\sqrt{n-1}}$

$SE(R)$ (the estimate) $= SD(R)$ (the estimator) $= \sigma_R \approx s_R$

7.8 Ratio Estimates

Suppose we know $x_1, \ldots, x_N$ and we want to estimate μ_y — take a random sample $Y_1, \ldots, Y_N$ (that we can match to x)

Ratio estimate of μ_y, $\bar{Y}_R = \frac{\mu_x}{\bar{X}}\bar{Y} = \mu_x R$

- Motivation: there might be correlation between x and y, so we can adjust for that to give a better estimate

R estimates $r = \frac{\mu_y}{\mu_x}$ well when n is large, so we expect $\bar{Y}_R$ to estimate $\mu_x r = \mu_y$ well too; is it 'better' than just using $\bar{Y}$?

Expectation and Variance of Ratio Estimate

$Var(\bar{Y}_R) \approx \frac{1}{n}(\frac{N-n}{N-1})(r^2\sigma_x^2 + \sigma_y^2 - 2r\rho\sigma_x\sigma_y)$

$E(\bar{Y}_R) \approx \mu_y + \frac{1}{n}(\frac{N-n}{N-1})\frac{1}{\mu_x}(r\sigma_x^2 - \rho\sigma_x\sigma_y)$

Both $\bar{Y}_R$ and $\bar{Y}$ are roughly unbiased when n is large.

But $\bar{Y}_R$ can have a smaller variance, which makes for a better estimator. How do we show this?

Let $C_x = \frac{\sigma_x}{\mu_x}$ and $C_y = \frac{\sigma_y}{\mu_y}$ (coefficient of variation).

$\bar{Y}_R$ is a better estimator than $\bar{Y}$ of μ_y if $\rho > \frac{1}{2}(\frac{C_x}{C_y})$:

$$\begin{aligned}
Var(\bar{Y}_R) &\approx \frac{1}{n}(\frac{N-n}{N-1})(r^2\sigma_x^2 + \sigma_y^2 - 2r\rho\sigma_x\sigma_y) \\
&< \frac{1}{n}(\frac{N-n}{N-1})(r^2\sigma_x^2 + \sigma_y^2 - 2r(\frac{1}{2})(\frac{\sigma_x/\mu_x}{\sigma_y/\mu_y})\sigma_x\sigma_y) \\
&= \frac{1}{n}(\frac{N-n}{N-1})(r^2\sigma_x^2 + \sigma_y^2 - r^2\sigma_x^2) \\
&= \frac{1}{n}(\frac{N-n}{N-1})\sigma_y^2 \\
&= Var(\bar{Y})
\end{aligned}$$

(???) What about negative correlation? Won't that still give some useful information that makes $\bar{Y}_R$ a better estimator than $\bar{Y}$?

Estimating Variance, Constructing CI for Ratio Estimate

Estimating $Var(\bar{Y}_R)$: $s^2_{\bar{Y}_R} = \frac{1}{n}(\frac{N-n}{N-1})(R^2 s_x^2 + s_y^2 - 2Rs_{xy})$

Approximate $(1-\alpha)$ CI for μ_y: $\bar{Y}_R \pm z_{\frac{\alpha}{2}} s_{\bar{Y}_R}$

8 Parameter Estimation A: Introduction, Method of Moments

How do we estimate parameters of a population distribution, more generally?

8.1 Example: Radioactive Emission

Numbers of emissions are given to have Poisson distritubitions, which have the same rate and are independent.

Data: $n = 1207$ realisations of $X_1, \ldots, X_n$ where each is $Po(\lambda)$

- Expected column for $n = k$ is computed from $1207 \times \frac{\lambda^k e^{-\lambda}}{k!}$, where $\lambda \approx 8.392$ (we got this estimate for λ by survey sampling, see this later)

n	Observed	Expected
0-2	18	12.2
3	28	27.0
4	56	56.5
5	105	94.9
6	126	132.7
7	146	159.1
8	164	166.9
9	161	155.6
10	123	130.6
11	101	99.7
12	74	69.7
13	53	45.0
14	23	27.0
15	15	15.1
16	9	7.9
17+	5	7.1

Goal: estimate λ using experiment results

Refresher on Poisson: $P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$, $E(X) = Var(X) = \lambda$

Estimating λ

(Borrowing from survey sampling)

- $\bar{x} = 8.392$ is an estimate of $\mu = \lambda$
- $SE(\bar{x}) = \sqrt{\frac{\lambda}{n}}$, approx $SE = 0.08$

8.2 General Estimation Problem

Let $X_1, \ldots, X_n$ be IID random variables with density $f(x|\theta)$, where $\theta \in \mathbb{R}^p$ is an unknown constant

Realisations $x_1, \ldots, x_n$ used to get an estimate for θ, which is a realisation of random variable estimator $\hat{\theta}$

- $Bias = E(\hat{\theta}) - \theta$
- $SE = SD(\hat{\theta})$

Moments

Let $\mu_k = E(X^k)$, the k-th moment of X

- $Var(X) = \mu_2 - \mu_1^2$
- For $Pois(\lambda)$, $\mu_1 = \lambda$, $\mu_2 = \lambda + \lambda^2$

- For $N(\mu, \sigma^2)$, $\mu_1 = \mu$, $\mu_2 = \sigma^2 + \mu^2$

Estimating Moments

k-th sample moment $\hat{\mu}_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$, which is an estimator of the k-th moment μ_k

- Realisation of $\hat{\mu}_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k$

8.3 The Method of Moments (MOM)

Express your estimator $\hat{\theta} = g(\hat{\mu}_1, \hat{\mu}_2, \ldots)$ (a function of the moments)

Example of MOM: Poisson

Let $X_i \sim Po(\lambda)$.

MOM estimator for λ: $\hat{\lambda} = \hat{\mu}_1 = \bar{X}$

- Then realisation of $\bar{X}$ is $\bar{x} = 8.392$
- Then MOM estimate of λ is 8.392, where approx SE is 0.08
- $Bias = E(\bar{X}) - \mu = 0$, $SE = SD(\bar{X}) = \sqrt{\frac{\lambda}{n}} \approx \sqrt{\frac{\bar{x}}{n}}$

Further, we can have normal approximation ($\bar{X}$ is approximately normal because of Poisson approximation):

- Approximate 95%-CI for λ is $8.932 \pm 1.96 \times 0.08 = (8.24, 8.55)$

Example of MOM: Normal

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

$\theta = (\mu, \sigma^2)$ is an unknown vector.

We know these things:

- $\mu_1 = \mu$
- $\mu_2 = \sigma^2 + \mu^2$

So we can express MOM estimators $\hat{\mu}$ and $\hat{\sigma}^2$ in terms of X's:

- $\hat{\mu} = \hat{\mu}_1 = \bar{X}$
- $\hat{\sigma}^2 = \hat{\mu}_2 - \hat{\mu}_1^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - (\bar{X})^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$ (since $\sigma^2 = \mu_2 - \mu_1^2$)

Questions

Joint distribution of $(\hat{\mu}, \hat{\sigma}^2)$?

- $\hat{\mu}$ and $\hat{\sigma}^2$ are independent, so we only need to look at marginal distributions; joint distribution is just product of marginal distributions
- $(\star)$ $\hat{\mu} \sim N(\mu, \frac{\sigma^2}{n})$ — since $\hat{\mu}$ is $\bar{X}$
- $(\star)$ $\frac{n\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-1}$ i.e. $\hat{\sigma}^2 = \frac{\sigma^2}{n} U_{n-1}$ where $U_{n-1} \sim \chi^2_{n-1}$ — we proved this earlier in chapter 2

Are the estimators biased?

- $E(\hat{\mu}) = \mu$
- $E(\hat{\sigma^2}) = \frac{n-1}{n}\sigma^2 = \sigma^2 - \frac{\sigma^2}{n}$ (biased)

What are the SEs? How can they be estimated and interpreted?

- $SE(\hat{\mu})$ (estimate) $= SD(\hat{\mu})$ (estimator) $= \frac{\sigma}{\sqrt{n}} \approx \frac{s}{\sqrt{n}}$ or $\frac{\hat{\sigma}}{\sqrt{n}}$
- $SE(\hat{\sigma}^2)$ (estimate) $= SD(\hat{\sigma}^2)$ (estimator) $= \sqrt{Var(\hat{\sigma}^2)}$ (see below)
- $Var(\hat{\sigma}^2) = \frac{\sigma^4}{n^2}(2(n-1)) = \frac{2}{n}(\frac{n-1}{n})\sigma^4 \approx \frac{2}{n}(\frac{n-1}{n})s^4$ (or replace with $\hat{\sigma}^4$)

Example of MOM: Gamma

Gamma distribution with shape α and rate λ has density function

$f(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}$, $x > 0$

$\theta = (\lambda, \alpha)$ is an unknown vector.

We know these things:

- $\mu_1 = \frac{\alpha}{\lambda}$
- $\mu_2 = \frac{\alpha(\alpha+1)}{\lambda^2}$

So we can express MOM estimators $\hat{\lambda}$ and $\hat{\alpha}$ in terms of X's:

- $\hat{\lambda} = \frac{\hat{\mu}_1}{\hat{\mu}_2 - \hat{\mu}_1^2} = \frac{\bar{X}}{\hat{\sigma}^2}$ (which can be estimated with $\frac{\bar{x}}{\hat{\sigma}^2}$)
- $\hat{\alpha} = \frac{\hat{\mu}_1^2}{\hat{\mu}_2 - \hat{\mu}_1^2} = \frac{\bar{X}^2}{\hat{\sigma}^2}$ (which can be estimated with $\frac{\bar{x}^2}{\hat{\sigma}^2}$)

(Working)

$$\begin{aligned} \mu_2 &= \frac{\alpha^2}{\lambda^2} + \frac{\alpha}{\lambda}(\frac{1}{\lambda}) \\ &= \mu_1^2 + \mu_1(\frac{1}{\lambda}) \\ \mu_2 - \mu_1^2 &= \mu_1(\frac{1}{\lambda}) \\ \Rightarrow \ \lambda &= \frac{\mu_1}{\mu_2 - \mu_1^2} \\ \Rightarrow \ \alpha &= \lambda\mu_1 = \frac{\mu_1^2}{\mu_2 - \mu_1^2} \end{aligned}$$

Bias and SE?

- Bias of $\hat{\lambda}$ estimate: $E_{\lambda,\alpha}(\hat{\lambda}) - \lambda = ???$
- SE of $\hat{\lambda}$ estimate: $SD_{\lambda,\alpha}(\hat{\lambda}) = ???$
- Same goes for bias and SE of $\hat{\alpha}$ estimate — no nice analytical expression!

Rainfall example (Rice p.264)

Assume samples are realisations of IID Gamma RVs $X_1, \ldots, X_{227}$ with unknown shape α and rate λ

$n = 227$, sample mean $\bar{x} = 0.224$, sample variance $\hat{\sigma}^2 = 0.1338$

Then MOM estimates of λ and α are:

- $\hat{\lambda}$ (estimate) $= \frac{0.224}{0.1338} \approx 1.67$
- $\hat{\alpha}$ (estimate) $= \frac{0.224^2}{0.1338} \approx 0.38$

Example of MOM: Angular Distribution

Let $X = \cos\theta$, where θ is random angle of electron emission

$f(x) = \frac{1+\alpha x}{2}$, $x \in [-1, 1]$ where $\alpha \in [-1, 1]$ is an unknown constant

Find the MOM estimator for α, based on IID $X_1, \ldots, X_n$:

- $E(X) = \int_{-1}^{1} x \cdot \frac{1+\alpha x}{2}\, dx = \int_{-1}^{1} \frac{x}{2} + \frac{\alpha}{2}x^2\, dx = [\frac{x^2}{4} + \frac{\alpha}{6}x^3]_{-1}^{1} = \frac{\alpha}{3}$
- $\mu = \frac{\alpha}{3}$, so $\alpha = 3\mu$, hence $\hat{\alpha} = 3\hat{\mu}_1 = 3\bar{X}$

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

MOM estimators are generally consistent (proof omitted).

Summary of MOM

- Method of moments: first, we express (population) moments in terms of parameters of interest; next, we express the parameters in terms of the moments. Then MOM estimator of the parameter is obtained by estimating moments with sample moments.
- MOM estimators can be biased or unbiased, but are generally consistent (asymptotically unbiased, as $n \to \infty$ then the estimator converges to the parameter)

9 Parameter Estimation B: Bootstrap and Monte Carlo, Parametric Family of Distributions

Combining bootstrap + monte carlo is powerful tool to estimate biases and SEs

9.1 Bootstrap

Bootstrap: to approximate the error of our estimate, we use the estimate itself (plug in estimate in its estimate)

- E.g. Poisson: the MOM estimate $\hat{\lambda} = \bar{x}$ has $SE(\bar{x}) = SD(\bar{X}) = \sqrt{\frac{\lambda}{n}} \approx \sqrt{\frac{\bar{x}}{n}}$

Problem with bootstrap: sometimes, the quantities we want can't be simply expressed in terms of unknown parameters

- E.g. Gamma: the MOM estimate $\hat{\lambda} = \frac{\bar{X}}{\hat{\sigma}^2}$ has expectation and variance that's difficult to compute, even if we know λ and α

9.2 Monte Carlo

Monte Carlo: use a large sample to approximate an expectation

- Justified by LLN
- Useful to estimate expectations when there isn't a nice analytical form for it
- Very powerful if we know the underlying true distribution, and so can repeat the simulation many times: then estimate will be very close to expectation

Example: Gamma Distribution

(For the previous problem with $\theta = (\lambda, \alpha)$)

If we know $\theta = (\lambda, \alpha)$, then we can simulate 227 realisations from $Gamma(\lambda, \alpha)$ to obtain a realisation of MOM estimator $\hat{\lambda}$. Repeat this say 10,000 times to obtain 10,000 realisations of $\hat{\lambda}$.

By LLN, expectation and SD of $\hat{\lambda}$ is approximately the average and SD of 10,000 realisations of $\hat{\lambda}$ — so we just approximated its expectation and SD!

9.3 Combo: Bootstrap then Monte Carlo

Example: Gamma Distribution

Previously, we obtained estimate of $\hat{\lambda} = 1.67$, estimate of $\hat{\alpha} = 0.38$. How to find bias and SE of $\hat{\lambda}$ and $\hat{\alpha}$ estimates?

Bootstrap step

Key insight: my world with unknown (λ, α) behaves very similarly to a world where truth is the estimate $(1.67, 0.38)$, so bootstrap.

- Suppose we conjure a different world where some data is generated from $Gamma(1.67, 0.38)$.
- Suppose in that world, we don't know its λ or α (which are actually 1.67 and 0.38), and want to estimate it by MOM; estimators are $1.\hat{6}7$ and $0.\hat{3}8$
- Back in our world, since the sample size is large, $(1.67, 0.38)$ is likely close to (λ, α), so in distribution, $(\hat{\lambda} - \lambda, \hat{\alpha} - \alpha) \approx (1.\hat{6}7 - 1.67, 0.\hat{3}8 - 0.38)$ — this is the bootstrap
- i.e. what we have ($\hat{\lambda}$ and $\hat{\alpha}$) is similar to the "ground truth"; our world is close to the other world by bootstrapping

Hence for λ,

- $bias(1.67) = E_{\lambda,\alpha}(\hat{\lambda}) - \lambda \approx E_{1.67,0.38}(\hat{1.67}) - 1.67$
- $SE(1.67) = SD_{\lambda,\alpha}(\hat{\alpha}) \approx SD_{1.67,0.38}(\hat{1.67})$
- and similarly for α.

Monte Carlo step

- Generate 10,000 realisations for $\hat{1.67}$, each time using 227 samples from the $Gamma(1.67, 0.38)$ distribution
- $E_{1.67,0.38}(\hat{1.67}) - 1.67 \approx 0.09$, $SD_{1.67,0.38}(\hat{1.67}) \approx 0.35$
- $E_{1.67,0.38}(\hat{0.38}) - 0.38 \approx 0.02$, $SD_{1.67,0.38}(\hat{0.38}) \approx 0.06$

Hence $bias(1.67) \approx 0.09$, $SE(1.67) \approx 0.35$); $bias(0.38) \approx 0.02$, $SE(0.38) \approx 0.06$

Summary

First, we obtain a realisation of $\hat{\lambda}$, which is 1.67. Then we approximate:

- $bias(1.67) = E_{\lambda,\alpha}(\hat{\lambda}) - \lambda \approx E_{1.67,0.38}(\hat{1.67}) - 1.67$ (by bootstrap) ≈ 0.09 (by Monte Carlo)
- $SE(1.67) = SD_{\lambda,\alpha}(\hat{\alpha}) \approx SD_{1.67,0.38}(\hat{0.38})$ (by bootstrap) ≈ 0.35 (by Monte Carlo)

Similarly for α:

- $bias(0.38) = \ldots \approx \ldots \approx 0.02$
- $SE(0.38) = \ldots \approx \ldots \approx 0.06$

Hence we estimate λ to be 1.58 ± 0.35, and estimate α to be 0.36 ± 0.06.

9.4 Parametric Family of Distributions

Idea: think of not a specific distribution, but a whole set of distributions

- e.g. Poisson family of distributions: $\{Poisson(\lambda) \mid \lambda \in \mathbb{R}, \lambda > 0\}$

Let $\{f(x|\theta) \mid \theta \in \Theta \subset \mathbb{R}^p\}$ be a parametric family, where Θ is the parameter space.

Let $x_1, \ldots, x_n$ be realisations of IID RVs $X_1, \ldots, X_n$ with density $f(x|\theta_0)$, where $\theta_0 \in \Theta$ is an unknown parameter we want to estimate from the data.

- Poisson family: $\theta = \lambda$, $\Theta = (0, \infty)$, $f(x|\theta) = \frac{\lambda^x e^{-\lambda}}{x!}$, $x = 0, 1, \ldots$
- Normal family: $\theta = (\mu, \sigma^2)$, $\Theta = (-\infty, \infty) \times (0, \infty)$, $f(x|\theta) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$, $-\infty < x < \infty$
- Gamma family: $\theta = (\alpha, \lambda)$, $\Theta = (0, \infty)^2$, $f(x|\theta) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}$, $x > 0$

Assumption of *identifiability*: function $\theta \to f(\cdot|\theta)$ is one-one, i.e. there cannot be $\theta_1 \neq \theta_2$ such that they give the same distribution $f(\cdot|\theta_1) = f(\cdot|\theta_2)$

- Each distribution can be mapped back to a unique θ, a.k.a. no two different θ's lead to the same distribution

9.5 Mean Square Error

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

10 Parameter Estimation C: Maximum Likelihood

ML, like MOM, yields consistent estimators.

ML estimates are asymptotically the most efficient among consistent estimates, and has the smallest SE $\Rightarrow$ better than MOM in a way

(Summary)

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

10.1 Likelihood Function

Suppose data come from density specified by a general θ. Then $P(X_1 = x_1, \ldots, X_n = x_n) = \prod_{i=1}^{n} f(x_i|\theta)$.

Likelihood function is $\theta \to L(\theta) = \prod_{i=1}^{n} f(x_i|\theta)$.

ML estimate of θ_0 is the number that maximises the likelihood over Θ.

(Both functions are mathematically the same, but the intepretation is different. Density function is seen as function of outcomes; likelihood function is seen as function of parameters)

10.2 Maximum Likelihood Estimator

Random likelihood function: $L(\theta) = \prod_{i=1}^{n} f(X_i|\theta)$

Maximum likelihood estimator $\hat{\theta}_0$: find it by maximising the random likelihood function

- $bias = E_{\theta_0}(\hat{\theta}_0) - \theta_0$
- $SE = SD_{\theta_0}(\hat{\theta}_0)$

Loglikelihood function: apply logarithm on likelihood function

10.3 Example: Poisson distribution

Let $x_1, \ldots, x_n$ be realisations from IID $Po(\lambda_0)$ RVs, where $\lambda_0 > 0$ is an unknown parameter

Likelihood function is $L(\lambda) = \prod_{i=1}^{n} f(x_i|\lambda_0) = \prod_{i=1}^{n} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} = \frac{\lambda^{\sum_{i=1}^{n} x_i} e^{-n\lambda}}{\prod_{i=1}^{n} x_i!}$

Loglikelihood function is $\ell(\lambda) = \sum_{i=1}^{n} x_i \log \lambda - n\lambda - \sum_{i=1}^{n} \log x_i!$

- 1st derivative: $\ell'(\lambda) = \sum_{i=1}^{n} x_i(\frac{1}{\lambda}) - n$
- 2nd derivative: $\ell''(\lambda) = \sum_{i=1}^{n} x_i(-\frac{1}{\lambda^2}) < 0$ since $x_i > 0$ in Poisson distribution
- So max likelihood $\ell'(\hat{\lambda}_0) = 0$
- So max likelihood when $\frac{1}{\hat{\lambda}_0} \sum_{i=1}^{n} x_i - n = 0$
- So max likelihood when $\hat{\lambda}_0 = \bar{x}$

ML estimate of λ_0 is $\bar{x}$, so ML estimator is $\hat{\lambda}_0$ is $\bar{X}$.

10.4 Example: Normal distribution

Let $X_1, \ldots, X_n$ be IID $N(\mu, \sigma^2)$, where μ and σ are unknown parameters

- Note: symbols do not have subscript 0—either it means unknown parameter, or generic element of parameter set

Logdensity of $N(\mu, \sigma^2)$ is $\log f(x|\mu, \sigma) = -\log \sigma - \frac{\log 2\pi}{2} - \frac{(x-\mu)^2}{2\sigma^2}$

Random loglikelihood function is $\ell(\mu, \sigma) = \sum_{i=1}^n \log f(X_i|\mu, \sigma) = -n \log \sigma - \frac{n \log 2\pi}{2} - \frac{\sum_{i=1}^n (X_i-\mu)^2}{2\sigma^2}$

$$\frac{\partial l}{\partial \mu} = -\frac{1}{\sigma^2} \sum_i (X_i - \mu)(-1) = \frac{1}{\sigma^2}(\sum_i X_i - n\mu)$$
$$\frac{\partial l}{\partial \sigma} = -\frac{n}{\sigma} + \frac{\sum_i (X_i - \mu)^2}{\sigma^3} = \frac{1}{\sigma}\left(\frac{\sum_i (X_i - \mu)}{\sigma^2} - n\right)$$

$$\frac{\partial l}{\partial \mu}(\hat{\mu}, \hat{\sigma}) = 0$$
$$\frac{1}{\hat{\sigma}^2}(\sum_i X_i - n\mu) = 0$$
$$\Rightarrow \hat{\mu} = \frac{1}{n} \sum_i X_i = \bar{X}$$

$$\frac{\partial l}{\partial \sigma}(\hat{\mu}, \hat{\sigma}) = 0$$
$$\frac{\sum_i (X_i - \hat{\mu})}{\hat{\sigma}^2} - n = 0$$
$$\Rightarrow \hat{\sigma} = \frac{1}{n} \sum_i (X_i - \bar{X})^2$$

Note that the MLE of σ^2 is $\hat{\sigma}^2$, so in a way it's better than s^2 even though $\hat{\sigma}^2$ has some bias

10.5 Example: Gamma distribution

Let $X_1, \ldots, X_n$ be IID $Gamma(\lambda, \alpha)$, where λ and α are unknown parameters

Logdensity of $Gamma(\lambda, \alpha)$ is $\alpha \log \lambda + (\alpha - 1) \log x - \lambda x - \log \Gamma(\alpha)$

Random loglikelihood function is $\ell(\lambda, \alpha) = n\alpha \log \lambda + (\alpha - 1) \sum_{i=1}^n \log X_i - \lambda \sum_{i=1}^n X_i - n \log \Gamma(\alpha)$

$$\frac{\partial l}{\partial \lambda} = \frac{n\alpha}{\lambda} - \sum_i X_i$$
$$\frac{\partial l}{\partial \alpha} = n \log \lambda + \sum_i \log X_i - n \frac{\Gamma'(\alpha)}{\Gamma(\alpha)}$$

The ML estimator $\hat{\theta} = (\hat{\lambda}, \hat{\alpha})$ satisfies $\frac{\partial l}{\partial \lambda}(\hat{\mu}, \hat{\sigma}) = 0$ and $\frac{\partial l}{\partial \alpha}(\hat{\mu}, \hat{\sigma}) = 0$, so:

$$n \log \hat{\alpha} - n \log \bar{X} + \sum_{i=1}^n \log X_i - n \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} = 0$$

$$\hat{\lambda} = \frac{\hat{\alpha}}{\bar{X}}$$

No closed analytical form, so we use numerical methods (e.g. Newton-Raphson) instead: find an $\hat{\alpha}$ that solves the equation to find MLE.

Here, MOM $\neq$ MLE!

Bias and SE of Gamma MLE

- Use Bootstrap + Monte Carlo on $Gamma(1.96, 0.44)$ again
- $Bias(1.96) = E_{\lambda,\alpha}(\hat{\lambda}) - \lambda \approx E_{1.96,0.44}(1.\hat{9}6) - 1.96 \approx 0.04$
- $SE(1.96) = SD_{\lambda,\alpha}(\hat{\lambda}) \approx SD_{1.96,0.44}(1.\hat{9}6) \approx 0.26$
- $Bias(0.44) = \ldots \approx 0.00$
- $SE(0.44) = \ldots \approx 0.03$
- So $\hat{\lambda} \approx 1.92 \pm 0.26$, $\hat{\alpha} \approx 0.44 \pm 0.03$

10.6 Example: Multinomial distribution

Experiment has r outcomes, with probabilites $\mathbf{p} = (p_1, \ldots, p_r)$

- Similar to r sided die, each side with different probability, roll n times

Let X_i be the number of times that outcome i occurs in n independent runs of the experiment.

$(X_1, \ldots, X_r)$ has multinomial distribution, with density:

$f(x_1, \ldots, x_r) = \binom{n}{x_1 \ldots x_r} \prod_{i=1}^{r} p_i^{x_i}$

Properties

- $E(X_i) = np_i$
- $Var(X_i) = np_i(1 - p_i)$
- $Cov(X_i, X_j) = -np_ip_j$ for $i \neq j$

Estimating $\mathbf{p}$ with ML

- Let $(x_1, \ldots, x_n)$ be a realisation of $(X_1, \ldots, X_r) \sim Multinomial(n, \mathbf{p})$
- Loglikelihood function $\ell(\mathbf{p}) = \kappa + \sum_{i=1}^{r} x_i \log p_i$, where $\kappa = \log \binom{n}{x_1 \ldots x_r}$ does not depend on $\mathbf{p}$
- Note that since the total probability equals to 1 where $p_r = 1 - p_1 - \ldots - p_{r-1}$, there are only $r - 1$ free variables

$$\begin{aligned}
\frac{\partial \ell}{\partial p_i} &= 0 + \left(0 + \ldots + \frac{x_i}{p_i} + 0 + \ldots + \frac{x_r}{1 - p_1 - \ldots - p_{r-1}}(-1)\right) \\
&= \frac{x_i}{p_i} - \frac{x_r}{1 - p_1 - \ldots - p_{r-1}} \\
&= \frac{x_i}{p_i} - \frac{x_r}{p_r}
\end{aligned}$$

Letting $\frac{\partial \ell}{\partial p_i} = 0$:

$$\frac{x_i}{p_i} - \frac{x_r}{p_r} = 0$$
$$p_r x_i = p_i x_r$$
$$\sum_{i=1}^{r} p_r x_i = \sum_{i=1}^{r} x_r p_i$$
$$p_r \sum_{i=1}^{r} x_i = x_r \sum_{i=1}^{r} p_i$$

$$p_r = \frac{x_r}{n}$$
$$\Rightarrow (\frac{x_r}{n}) x_i = x_r p_i$$
$$\Rightarrow p_i = \frac{x_i}{n}$$

Hence ML estimate of $\hat{\mathbf{p}}$:

- ML estimator $\hat{p_i} = \frac{X_i}{n}$
- ML estimate of $\hat{p_i} = \frac{x_i}{n}$

Bias and variance of $\hat{\mathbf{p}}_i$:

- $E(\hat{p_i}) = \frac{E(X_i)}{n}$
- $Var(\hat{p_i}) = \frac{np_i(1-p_i)}{n^2} = \frac{p_i(1-p_i)}{n}$
- $Cov(\hat{p_i}, \hat{p_j}) = \frac{1}{n^2} Cov(X_i, X_j) = -\frac{p_i p_j}{n}$

MOM estimate of $\mathbf{p}$:

- We know $E(X_i) = np_i$, so MOM $\hat{p_i} = \frac{X_i}{n}$

10.7 Example: HWE Trinomial (related to Multinomial)

Hardy-Weinberg equilibrium

- Suppose there are only 2 alleles A and a, where proportion of a in population is θ
- Assume that population is very large and mating is completely random
- Then genotype proportions of the next generation are:
  - $AA : (1-\theta)^2$
  - $Aa : 2\theta(1-\theta)$
  - $aa : \theta^2$

Under HWE, number of a alleles in child is the sum of two independent $Ber(\theta)$ random variables, i.e. $Bin(2, \theta)$

- Hence the number of a alleles in n children is $Bin(2n, \theta)$

Setup of example problem:

- Suppose sample frequencies are as such: $AA : x_1 = 342$, $Aa : x_2 = 500$, $aa : x_3 = 187$ — $n = 1029$
- Frequencies are approximately realisations from $(X_1, X_2, X_3) \sim Multinomial(1029, \mathbf{p})$ for some $\mathbf{p}$
- Assuming HWE holds, $\mathbf{p} = ((1-\theta)^2, 2\theta(1-\theta), \theta^2)$, where θ is an unknown constant

MOM estimators of θ

$$E(X_1) = np_1 = n(1-\theta)^2$$
$$\Rightarrow \theta = 1 - \sqrt{\frac{E(X_1)}{n}}$$
$$E(X_3) = np_3 = n\theta^2$$
$$\Rightarrow \theta = \sqrt{\frac{E(X_3)}{n}}$$

Some possible MOM estimators of θ:

- MOM estimator 1 is $\hat{\theta} = 1 - \sqrt{\frac{X_1}{n}}$
- MOM estimator 2 is ...
- MOM estimator 3 is $\hat{\theta} = \sqrt{\frac{X_3}{n}}$
- (Here, it's OK to use X_1 and X_3 instead of $\bar{X}_1$ and $\bar{X}_3$ because we have only 1 sample, $n = 1$)

ML estimator

$L(\theta) = \frac{n!}{x_1!x_2!x_3!}p_1^{x_1}p_2^{x_2}p_3^{x_3} = \frac{n!}{x_1!x_2!x_3!}(1-\theta)^{2x_1+x_2}\theta^{x_2+2x_3}2^{x_2}$

$\ell(\theta) = \kappa + (2x_1 + x_2)\log(1-\theta) + (x_2 + 2x_3)\log\theta$

$\ell'(\theta) = -\frac{2x_1+x_2}{1-\theta} + \frac{x_2+2x_3}{\theta}$

$\ell''(\theta) = \ldots < 0$ (confirming that it's a maximum)

$\ell'(\hat{\theta}) = 0 \Rightarrow -\frac{2x_1+x_3}{1-\hat{\theta}} + \frac{x_2+2x_3}{\hat{\theta}} = 0 \Rightarrow \ldots \Rightarrow \hat{\theta} = \frac{x_2+2x_3}{2n}$

ML estimate is $\hat{\theta} = \frac{x_2+2x_3}{2n} = \frac{500+2\times187}{2\times1029} \approx 0.42$

ML estimator is $\hat{\theta} = \frac{X_2+2X_3}{2n}$

- We realise that $X_2 + 2X_3$ is the number of a alleles, so $X_2 + 2X_3 \sim Bin(2n, \theta)$
- So $E(\hat{\theta}) = \frac{1}{2n}E(X_2 + 2X_3) = \frac{1}{2n}2n\theta = \theta$
- So $Var(\hat{\theta}) = \frac{1}{4n^2}Var(X_2 + 2X_3) = \frac{1}{4n^2}2n\theta(1-\theta) = \frac{\theta(1-\theta)}{2n}$
- So $SE(0.42) = SD(\hat{\theta}) = \sqrt{0.42 \times 0.58}2058 \approx 0.01$

10.8 Confidence Intervals based on MLE

MLEs are asymptotically normal, and have the smallest SE within the class of consistent estimators

Let $\theta \in \Theta$ be an unknown constant, with ML estimate $\hat{\theta}$

- If sample size is large, we can construct approximate CI for θ, relying on CLT

ML estimator of μ is $\bar{X}$

- We have $\frac{\bar{X}-\mu}{S-\sqrt{n}} \sim t_{n-1}$
- $P(-t_{n-1,\alpha/2}\frac{s}{\sqrt{n}} \leq \ldots \leq t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}) = 1 - \alpha$
- Given realisations $\bar{x}$ and s, an exact $(1-\alpha)$ CI for μ is $(\bar{x} - t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}, \bar{x} + t_{n-1,\alpha/2}\frac{s}{\sqrt{n}})$

ML estimator of σ^2 is $\hat{\sigma}^2$

- We have $\frac{n\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-1}$
- $P(\chi^2_{n-1,\alpha/2} \leq \ldots \leq \chi^2_{n-1,1-\alpha/2}) = 1 - \alpha$
- Given realisation $\hat{\sigma}^2$, an exact $(1-\alpha)$ CI for σ^2 is $(\frac{n\hat{\sigma}^2}{\chi^2_{n-1,\alpha/2}}, \frac{n\hat{\sigma}^2}{\chi^2_{n-1,1-\alpha/2}})$

ML estimators, in general, are asymptotically normally distributed, so can be used to construct CIs

11 Fisher Information

Recall the following things:

- Let $\hat{\theta}_0$ be the ML estimator of unknown parameter θ_0, based on n IID RVs $X_1, \ldots, X_n$ with density $f(x|\theta_0)$
- Bias of $\hat{\theta}_0$ is $E(\hat{\theta}_0) - \theta_0$
- SE of $\hat{\theta}_0$ is $SD(\hat{\theta}_0)$
- How to find approximate SD as $n \to \infty$? (Distribution of $\hat{\theta}_0$ becomes approximately normal)

Fisher information matrix at θ is a $p \times p$ matrix

- Given a parametric family of densities $\{f(x|\theta) \mid \theta \in \Theta \subset \mathbb{R}^p\}$
- $I(\theta) = -\int_{-\infty}^{\infty} \left[\frac{\partial^2}{\partial\theta^2}\{\log f(x|\theta)\}\right] f(x|\theta)\ dx$
- i.e. $I_{ij}(\theta) = -\int_{-\infty}^{\infty} \left[\frac{\partial^2}{\partial\theta_i\partial\theta_j}\{\log f(x|\theta)\}\right] f(x|\theta)\ dx$

Fisher information as *expectation*

- $I(\theta) = -E\left[\frac{\partial^2}{\partial\theta^2}\log f(X|\theta)\right]$, where X is a random varable with density $f(x|\theta)$
- Negative expectation of the [second derivative of the random log likelihood]

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)

11.1 Example: Poisson, $\theta = \lambda$

Poisson density: $f(x|\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}$, $x \geq 0$

Random log density: $\log f(X|\lambda) = X\log\lambda - \lambda - \log X!$

- $\frac{\partial}{\partial\lambda}\log f(X|\lambda) = \frac{X}{\lambda} - 1$
- $\frac{\partial^2}{\partial\lambda^2}\log f(X|\lambda) = -\frac{X}{\lambda^2}$
- $\therefore I(\lambda) = \frac{E(X)}{\lambda^2} = \frac{1}{\lambda}$
- Interpretation: Since $I(\lambda) = \frac{1}{\lambda}$, the larger λ is, the smaller $I(\lambda)$ is, so one sample of X gives less information on λ

11.2 Example: Bernoulli, $\theta = p$

Bernoulli density: $f(x|p) = p^x(1-p)^{1-x}$

Random log density: $\log f(X|p) = X\log p + (1-X)\log(1-p)$

- $\frac{\partial}{\partial p}\ldots = \frac{X}{p} - \frac{1-X}{1-p}$
- $\frac{\partial^2}{\partial p^2}\ldots = -\frac{X}{p^2} - \frac{1-X}{(1-p)^2}$
- $\therefore I(p) = \frac{E(X)}{p^2} + E(1-X)(1-p)^2 = \frac{1}{p} + \frac{1}{1-p} = \frac{1}{p(1-p)}$
- Interpretation: more information is given when p is close to 0 or 1; less information is given when p is close to $\frac{1}{2}$

11.3 Example: Normal, $\theta = (\mu, \sigma)$

Normal density: $f(x|\theta) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

Random log density: $\log f(X|\theta) = -\frac{\log 2\pi}{2} - \log\sigma - \frac{(X-\mu)^2}{2\sigma^2}$

- $\frac{\partial}{\partial\theta}\ldots = \left(\frac{X-\mu}{\sigma^2}, -\frac{1}{\sigma} + \frac{(X-\mu)^2}{\sigma^3}\right)$
- $\frac{\partial^2}{\partial\theta^2}\ldots = \begin{bmatrix} -\sigma^{-2} & -2\sigma^{-3}(X-\mu) \\ -2\sigma^{-3}(X-\mu) & \sigma^{-2} - 3\sigma^{-4}(X-\mu)^2 \end{bmatrix}$ (this is also called Hessian matrix, a square matrix of second-order partial derivatives)
- (NOTE) for $\frac{\partial^2}{\partial\theta_i\partial\theta_j}$, you can differentiate in either order and it'll give the same result!
- $\therefore I(\theta) = \begin{bmatrix} \sigma^{-2} & 2\sigma^{-3}E(X-\mu) \\ 2\sigma^{-3}E(X-\mu) & -\sigma^{-2} + 3\sigma^{-4}E(X-\mu)^2 \end{bmatrix} = \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{2}{\sigma^2} \end{bmatrix}$
- Interpretation: the larger the σ, the less information a sample gives us on both μ and σ

11.4 Example: Normal, $\theta = (\mu, v = \sigma^2)$

Normal density: $f(x|\theta) = \frac{1}{\sqrt{2\pi v}} e^{-\frac{(x-\mu)^2}{2v}}$

Random log density: $\log f(X|\theta) = -\frac{\log 2\pi}{2} - \frac{\log v}{2} - \frac{(X-\mu)^2}{2v}$

- $\frac{\partial}{\partial\theta}\ldots = \left(\frac{X-\mu}{v}, -\frac{1}{2v} + \frac{(X-\mu)^2}{2v^2}\right)$
- $\frac{\partial^2}{\partial\theta^2}\ldots = \begin{bmatrix} -\frac{1}{v} & -\frac{(X-\mu)}{v^2} \\ -\frac{(X-\mu)}{v^2} & \frac{1}{2v^2} - \frac{(X-\mu)^2}{v^3} \end{bmatrix}$
- $\therefore I(\theta) = \begin{bmatrix} \frac{1}{v} & 0 \\ 0 & \frac{1}{2v^2} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{bmatrix}$
- (NOTE) DIFFERENT from the one above, depending on the way we parameterise!

11.5 Example: Binomial, $\theta = p$

Random log density: $f(x|p) = \log\binom{n}{X} + X\log p + (n-X)\log(1-p)$

- $\frac{\partial}{\partial\theta}\ldots = 0 + \frac{X}{p} - \frac{n-X}{1-p}$
- $\frac{\partial^2}{\partial\theta^2}\ldots = -\frac{X}{p^2} - \frac{n-X}{(1-p)^2}$
- $\therefore I(\theta) = \frac{E(X)}{p^2} + \frac{n-E(X)}{(1-p)^2} = \ldots = \frac{n}{p(1-p)}$

Question: what's the difference in Fisher information between Binomial and n Bernoullis?

- No real difference
- Anyway, we see here that one sample from $Bin(n,p)$ is as informative as n IID samples from $Ber(p)$

11.6 Example: HWE Trinomial Distribution, θ

Let $\mathbf{X} = (X_1, X_2, X_3) \sim Multinomial(n, \mathbf{p})$ where $\mathbf{p} = ((1-\theta)^2, 2\theta(1-\theta), \theta^2)$

- $E(X_1) = n(1-\theta)^2$, $E(X_2) = 2n\theta(1-\theta)$, $E(X_3) = \theta^2$

Density function: $f(\mathbf{x}|\theta) = \ldots$

Random log density: $\log f(\mathbf{X}|\theta) = \kappa + (2X_1 + X_2)\log(1-\theta) + (X_2 + 2X_3)\log\theta$

- $\frac{\partial^2}{\partial\theta^2}\ldots = -\frac{2X_1+X_2}{(1-\theta)^2} - \frac{X_2+2X_3}{\theta^2}$
- $\therefore I(\theta) = \frac{2E(X_1)+E(X_2)}{(1-\theta)^2} + \frac{E(X_2)+2E(X_3)}{\theta^2} = \ldots = \frac{2n}{\theta(1-\theta)}$

11.7 Example: General Trinomial Distribution, $\theta = (p_1, p_2)$

Let $\mathbf{X} = (X_1, X_2, X_3) \sim Multinomial(n, \mathbf{p})$ where $p_3 = 1 - p_1 - p_2$

Multinomial density: $f(\mathbf{x}|\theta) = \frac{n!}{x_1!x_2!x_3!}p_1^{x_1}p_2^{x_2}p_3^{x_3}$

- $E(X_i) = np_i$

Random log density: $\log f(\mathbf{X}|\theta) = \kappa + X_1 \log p_1 + X_2 \log p_2 + X_3 \log(1 - p_1 - p_2)$

- $\frac{\partial}{\partial\theta}\ldots = (\frac{X_1}{p_1} - \frac{X_3}{p_3}, \frac{X_2}{p_2} - \frac{X_3}{p_3})$
- $\frac{\partial^2}{\partial\theta^2}\ldots = \begin{bmatrix} -\frac{X_1}{p_1^2} - \frac{X_3}{p_3^2} & -\frac{X_3}{p_3^2} \\ -\frac{X_3}{p_3^2} & -\frac{X_2}{p_2^2} - \frac{X_3}{p_3^2} \end{bmatrix}$
- $\therefore I(\theta) = n\begin{bmatrix} \frac{1}{p_1} + \frac{1}{p_3} & \frac{1}{p_3} \\ \frac{1}{p_3} & \frac{1}{p_2} + \frac{1}{p_3} \end{bmatrix}$

Trinomial Distribution With Only One Trial

Let $\mathbf{Y} = (Y_1, Y_2, Y_3) \sim Multinomial(1, \mathbf{p})$ where $p_3 = 1 - p_1 - p_2$

Multinomial density: $f(\mathbf{y}|\theta) = p_1^{y_1}p_2^{y_2}p_3^{y_3}$, where $\mathbf{Y} = (y_1, y_2, y_3) \in \{(1,0,0), (0,1,0), (0,0,1)\}$

11.8 Variance and Fisher Information

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

12 Large Sample Theory for MLE

All MLEs are *consistent* (asymptotically converges to what it's trying to estimate) and *asymptotically normal*; then we can form confidence intervals easily for MLEs

- Multivariate: we have $X_1, \ldots, X_n$; univariate: one sample of X
- θ can be a constant or a vector

12.1 ($\star$) Asymptotic Normality of MLE (θ as a Constant)

Let $X_1, \ldots, X_n$ be IID with density $f(\cdot|\theta)$, where θ is an unknown constant in $\Theta \subset \mathbb{R}$.

Let $\hat{\theta}$ be the MLE of θ. As $n \to \infty$, in distribution:

$$\sqrt{nI(\theta)}(\hat{\theta} - \theta) \to N(0, 1)$$

For large n, approximately:

$$\hat{\theta} \sim N\left(\theta, \frac{I(\theta)^{-1}}{n}\right)$$

In particular, as $n \to \infty$, $\hat{\theta} \to \theta$ — the MLE is consistent (asymptotically unbiased).

12.2 ($\star$) Asymptotic Normality of MLE (θ as a Vector)

Now let θ an unknown *vector* instead in $\Theta \subset \mathbb{R}^p$.

As $n \to \infty$, in distribution:

$$\sqrt{nI(\theta)}(\hat{\theta} - \theta) \to N(\mathbf{0}, \mathbf{I}_p)$$

For large n, approximately:

$$\hat{\theta} \sim N\left(\theta, \frac{I(\theta)^{-1}}{n}\right)$$

12.3 Interpretation

$nI(\theta)$ is amount of information in n IID samples with density $f(\cdot|\theta)$

- Asymptotic variance of MLE is inversely proportional to sample size n
- $I(\theta)^{-1}$ is similar to σ^2 in a sample survey

12.4 Example: Poisson

Let $X_1, \ldots, X_n$ be IID $Po(\lambda)$, where $\theta = \lambda$, $\hat{\theta} = \bar{X}$, $I(\theta) = \frac{1}{\lambda}$, so $I(\theta)^{-1} = \lambda$.

By the theorem, if n is large, then approximately $\bar{X} \sim N(\lambda, \frac{\lambda}{n})$.

12.5 Example: Normal Case (a) with $\theta = (\mu, \sigma)$

Let $X_1, \ldots, X_n$ be IID $N(\mu, \sigma^2)$, where $\theta = (\mu, \sigma)$, $\hat{\theta} = (\bar{X}, \hat{\sigma})$, $I(\theta) = \begin{bmatrix} 1/\sigma^2 & 0 \\ 0 & 2/\sigma^2 \end{bmatrix}$.

By the theorem, if n is large, then approximately $\begin{bmatrix} \bar{X} \\ \hat{\sigma} \end{bmatrix} \sim N\left(\begin{bmatrix} \mu \\ \sigma \end{bmatrix}, \begin{bmatrix} \sigma^2/n & 0 \\ 0 & \sigma^2/(2n) \end{bmatrix} \right)$ — bivariate normal distribution

- We already know that $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$ and that $\bar{X}$ and $\hat{\sigma}$ are independent
- But now we also know that $\hat{\sigma} \sim N(\sigma, \frac{\sigma^2}{2n})$
- (Note that $Cov(X_1, X_2) = 0$ does not imply that X_1 and X_2 are independent, but this is true of normal distributions)

12.6 Example: Normal Case (a) with $\theta = (\mu, v = \sigma^2)$

Let $X_1, \ldots, X_n$ be IID $N(\mu, \sigma^2)$, where $\theta = (\mu, v = \sigma^2)$, $\hat{\theta} = (\bar{X}, \hat{\sigma}^2)$, $I(\theta) = \begin{bmatrix} 1/\sigma^2 & 0 \\ 0 & 1/(2\sigma^4) \end{bmatrix}$.

By the theorem, if n is large, then approximately $\begin{bmatrix} \bar{X} \\ \hat{\sigma}^2 \end{bmatrix} \sim N\left(\begin{bmatrix} \mu \\ \sigma^2 \end{bmatrix}, \begin{bmatrix} \sigma^2/n & 0 \\ 0 & 2\sigma^4/n \end{bmatrix} \right)$

- We already know that $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$ and that $\bar{X}$ and $\hat{\sigma}$ are independent
- But now we also know that $\hat{\sigma}^2 \sim N(\sigma^2, \frac{2\sigma^4}{n})$ — which makes sense, since $\hat{\sigma}^2$ follows a χ^2 distribution

12.7 Example: HWE Trinomial

Let $\mathbf{W}_1, \ldots, \mathbf{W}_n$ be IID $Multinomial(1, \mathbf{p})$, where $\mathbf{p} = ((1-\theta)^2, 2\theta(1-\theta), \theta^2))$. $\mathbf{W}_i$ takes values (1,0,0), (0,1,0), (0,0,1) with probabilities as in $\mathbf{p}$.

Let $\mathbf{X} = \mathbf{W}_1 + \ldots + \mathbf{W}_n \sim Multinomial(n, \mathbf{p})$.

- Random likelihood: $L(\theta) = \prod_{i=1}^{n} (p_1^{\mathbf{W}_{i,1}} p_2^{\mathbf{W}_{i,2}} p_3^{\mathbf{W}_{i,3}})$
- Random loglikelihood: $\ell(\theta) = \sum_{i=1}^{n} (\mathbf{W}_{i,1} \log p_1 + \mathbf{W}_{i,2} \log p_2 + \mathbf{W}_{i,3} \log p_3) = \ldots = (2X_1 + X_2) \log(1-\theta) + (X_2 + 2X_3) \log \theta + X_2 \log 2$

So we find that MLEs based on the $\mathbf{W}$'s is the same as based on $\mathbf{X}$: $\hat{\theta} = \frac{X_2 + 2X_3}{2n}$

Fisher Information for HWE Trinomial

Let Fisher information based on $\mathbf{W}$'s be $I^*(\theta) = \frac{2}{\theta(1-\theta)}$.

- By our theorem, for large n, approximately $\hat{\theta} \sim N(\theta, \frac{\theta(1-\theta)}{2n})$

Let Fisher information based on $\mathbf{X}$ be $I(\theta) = nI^*(\theta)$.

- By our theorem, for large n, approximately $\hat{\theta} \sim N(\theta, I(\theta)^{-1})$
- It would be hard to apply the theorem directly on $\mathbf{X}$, since the sample size is 1 (similarly for 1 Binomial sample)

12.8 Example: General Trinomial

Let $\mathbf{W}_1, \ldots, \mathbf{W}_n$ be IID $Multinomial(1, \mathbf{p})$, where $\theta = (p_1, p_2)$.

Let $\mathbf{X} = \mathbf{W}_1 + \ldots + \mathbf{W}_n \sim Multinomial(n, \mathbf{p})$.

We also find that MLEs based on $\mathbf{W}$'s is the same as based on $\mathbf{X}$: $\hat{\theta} = (\frac{X_1}{n}, \frac{X_2}{n})$

Let Fisher information based on $\mathbf{W}$'s be $I^*(\theta) = \begin{bmatrix} 1/p_1 + 1/p_3 & 1/p_3 \\ 1/p_3 & 1/p_2 + 1/p_3 \end{bmatrix}$.

- For large n, approximately $\hat{\theta} = \begin{bmatrix} X_1/n \\ X_2/n \end{bmatrix} \sim N\left(\begin{bmatrix} p_1 \\ p_2 \end{bmatrix}, \begin{bmatrix} p_1(1-p_1)/n & -p_1p_2/n \\ -p_1p_2/n & p_2(1-p_2)/n \end{bmatrix} \right)$
- We already knew that expectation and variance are exact: $E(X_i) = np_i, Var(X_i) = np_i(1-p_i), Cov(X_i, X_j) = -np_ip_j$ (but now we're looking at X_i/n, so the constants involving n differ accordingly)
- But the approximate normality is new

12.9 SE and Bootstrap

Recall that SE of our estimate of θ is the SD of $\hat{\theta}$.

$$SE = SD(\hat{\theta}) \approx \sqrt{\frac{I(\theta)^{-1}}{n}}$$

(approximately for large n)

But now we have a problem: we don't actually know θ. So use the *bootstrap*, by calculating Fisher information at the estimate instead of θ:

$$SE \approx \sqrt{\frac{I(\hat{\theta})^{-1}}{n}}$$

(where $\hat{\theta}$ here denotes the *estimate* instead)

12.10 Random Intervals

Let $\hat{\theta}$ be the ML estimator for θ. For large n:

$$1 - \alpha \approx P\left(-z_{\alpha/2} \leq \frac{\hat{\theta} - \theta}{\sqrt{I(\theta)^{-1}/n}} \leq z_{\alpha/2} \right)$$

$$1 - \alpha \approx P\left(\hat{\theta} - z_{\alpha/2}\sqrt{\frac{I(\theta)^{-1}}{n}} \leq \theta \leq \hat{\theta} + z_{\alpha/2}\sqrt{\frac{I(\theta)^{-1}}{n}} \right)$$

We also estimate the approximate SE $\sqrt{\frac{I(\theta)^{-1}}{n}}$ with $\sqrt{\frac{I(\hat{\theta})^{-1}}{n}}$ (using bootstrap).

Hence the $(1-\alpha)$ CI for θ is approximately $\left(\hat{\theta} - z_{\alpha/2}\sqrt{\frac{I(\hat{\theta})^{-1}}{n}}, \hat{\theta} + z_{\alpha/2}\sqrt{\frac{I(\hat{\theta})^{-1}}{n}} \right)$ for large n.

Note that $SD(\hat{\theta})$ here is not exactly $\sqrt{\frac{I(\theta)^{-1}}{n}}$ (unlike previously), only approximately true with large n.

Three approximations, all good when n is large:

- Approximately normal (by MLE theorem)
- Approximate SD
- Approximate by bootstrap ($\hat{\theta}$ realisation is a good estimate for θ, consistency when n is large)

12.11 Example: CI for Poisson, $\theta = \lambda$

$\theta = \lambda$, estimated by $\hat{\theta} = \bar{x}$

$I(\theta)^{-1} = \lambda$, estimated by $I(\hat{\theta})^{-1} = \bar{x}$

For large n, approximate $(1-\alpha)$ CI for $\theta = \lambda$ is $\left(\bar{x} - z_{\alpha/2}\sqrt{\frac{\bar{x}}{n}}, \bar{x} + z_{\alpha/2}\sqrt{\frac{\bar{x}}{n}}\right)$

12.12 Example: CI for Normal a), $\theta = (\mu, \sigma)$

$\theta = (\mu, \sigma)$, estimated by $\hat{\theta} = (\bar{x}, \hat{\sigma})$

$I(\theta)^{-1} = \ldots$, estimated by $I(\hat{\theta})^{-1} = \begin{bmatrix} \hat{\sigma}^2 & 0 \\ 0 & \hat{\sigma}^2/2 \end{bmatrix}$

For large n, approximate $(1-\alpha)$ CI for μ is $\left(\bar{x} - z_{\alpha/2}\frac{\hat{\sigma}}{\sqrt{n}}, \bar{x} + z_{\alpha/2}\frac{\hat{\sigma}}{\sqrt{n}}\right)$

For large n, approximate $(1-\alpha)$ CI for σ is $\left(\hat{\sigma} - z_{\alpha/2}\frac{\hat{\sigma}}{\sqrt{2n}}, \hat{\sigma} + z_{\alpha/2}\frac{\hat{\sigma}}{\sqrt{2n}}\right)$

12.13 Example: CI for Normal a), $\theta = (\mu, \sigma^2)$

$\theta = (\mu, \sigma^2)$, estimated by $\hat{\theta} = (\bar{x}, \hat{\sigma}^2)$

$I(\theta)^{-1} = \ldots$, estimated by $I(\hat{\theta})^{-1} = \begin{bmatrix} \hat{\sigma}^2 & 0 \\ 0 & 2\hat{\sigma}^4 \end{bmatrix}$

For large n, approximate $(1-\alpha)$ CI for μ is $\left(\bar{x} - z_{\alpha/2}\frac{\hat{\sigma}}{\sqrt{n}}, \bar{x} + z_{\alpha/2}\frac{\hat{\sigma}}{\sqrt{n}}\right)$ (same as previous)

For large n, approximate $(1-\alpha)$ CI for σ is $\left(\hat{\sigma}^2 - z_{\alpha/2}\hat{\alpha}^2\sqrt{\frac{2}{n}}, \hat{\sigma} + z_{\alpha/2}\hat{\alpha}^2\sqrt{\frac{2}{n}}\right)$

12.14 Example: Bivariate Normal Distribution

$\mathbf{X} \sim N(\mu, \Sigma)$, where $\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$, $\Sigma = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$, $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$. $|\Sigma|$ is the determinant.

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)}$$

, where p is the number of parameters ($p = 2$ in the bivariate case)

Any bivariate normal $\mathbf{X}$ can be written as $\mathbf{X} = \mathbf{AZ} + \mathbf{b} \sim N(\mathbf{b}, \mathbf{AA}')$ (i.e. linear combination of standard normals)

12.15 Linear Regression

In linear regression (assuming independent normals), we minimize the squares because of MLEs!

Let $Y_1, \ldots, Y_n$ be RVs with $Y_i = \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \epsilon_i$ or $Y_i = \mu_i + \epsilon_i$, ie. $\mathbf{Y} = \boldsymbol{\mu} + \boldsymbol{\epsilon}$, where $\boldsymbol{\mu} = \boldsymbol{\beta X}$

- $\mathbf{X}$ is a fixed known $n \times p$ matrix
- $\boldsymbol{\beta}$ is a fixed unknown $p \times 1$ matrix
- $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I}_n)$ is a $n \times 1$ matrix, with fixed unknown σ^2
- $\mathbf{Y} \sim N(\mu, \sigma^2\mathbf{I}_n)$

Derivation: in the Multivariate Normal case, $\Sigma = \sigma^2 \boldsymbol{I}_n$

$$
\begin{aligned}
L(\theta) &= \frac{1}{(2\pi)^{\frac{p}{2}} |\sigma^2 \boldsymbol{I}_n|^{\frac{1}{2}}} \exp\{-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\sigma^2 \boldsymbol{I}_n)^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\} \\
&= \frac{1}{(2\pi)^{\frac{p}{2}} |\sigma^2 \boldsymbol{I}_n|^{\frac{1}{2}}} \exp\{-\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}')(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\} \\
\ell(\theta) &= -\frac{p}{2}\log 2\pi - n \log \sigma - \frac{1}{2\sigma^2}||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 \\
\frac{\partial}{\partial \sigma}\ell(\theta) &= -\frac{n}{\sigma} - \frac{1}{\sigma^3}||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2
\end{aligned}
$$

To find the MLE i.e. to maximise this log likelihood, we need to find $\boldsymbol{\beta}$ that minimizes the quantity $||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2$. Hence the MLE $\hat{\boldsymbol{\beta}}$ minimizes $||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2$, i.e. minimum squares (residual sum of squares)!

13 Efficiency

13.1 Cramer-Rao Inequality

($\star$) 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) $\frac{I(\theta)^{-1}}{n}$ tells you the best (i.e. lowest variance) you can ask from any unbiased estimator. No unbiased estimator can do better than this.

13.2 Efficiency and Relative Efficiency of *Unbiased* Estimators

Efficient unbiased estimators

- *Efficient*: an unbiased estimator $\hat{\theta}$ is *efficient* if $Var(\hat{\theta}) = \frac{I(\theta)^{-1}}{n}$ for every $\theta \in \Theta$
- Example of $X_1, \ldots, X_n$ IID Poisson: $\theta = \lambda$ estimated by $\bar{X}$ — $Var(\bar{X}) = \frac{\lambda}{n} = \frac{I(\theta)^{-1}}{n}$
- Example of $X_1, \ldots, X_n$ IID Bernoulli(p): $\theta = p$ estimated by $\hat{p}$ — $Var(\hat{p}) = \frac{p(1-p)}{n} = \frac{I(\theta)^{-1}}{n}$
- Example of $X_1, \ldots, X_n$ IID $N(\mu, \sigma^2)$: $\theta = (\mu, \sigma^2)$ estimated by $(\bar{X}, S^2)$ — $Var(\bar{X}) = \frac{\sigma^2}{n} = \frac{[I(\theta)^{-1}]_{11}}{n}$
  - But $Var(S^2) = \frac{2\sigma^4}{n-1} > \frac{[I(\theta)^{-1}]_{22}}{n} = \frac{2\sigma^4}{n}$
  - (S^2 is an unbiased estimator for σ^2, though note it's not from MLE)

Efficiency of unbiased estimators

- $Eff(\hat{\theta}) = \frac{I(\theta)^{-1}/n}{Var(\hat{\theta})} \leq 1$
- Example of S^2 in normal case: $Eff(S^2) = \frac{n-1}{n}$

Unbiased ML Estimators

In general, an ML estimator $\hat{\theta}$ is usually biased. But by asymptotic normality theorem, for large n it is approximately unbiased, and its variance is approximately $\frac{I(\theta)^{-1}}{n}$. Hence its efficiency is approximately 1 when n is large. Wonderful!

Relative efficiency of unbiased estimators

- $Eff(\tilde{\theta}, \hat{\theta}) = \frac{Var(\hat{\theta})}{Var(\tilde{\theta})}$ — allows for comparison without knowing the exact Fisher information
- $Eff(\tilde{\theta}, \hat{\theta}) = Eff(\tilde{\theta})$ if $\hat{\theta}$ is efficient

Relative Efficiency and Sample Sizes

The efficiency usually depends on n, and relative efficiencies also depend on n. Then strictly speaking we should be talking about $\tilde{\theta}_n$ and $\hat{\theta}_n$.

But a lot of the time, they have the form $\frac{k}{n}$, then the relative efficiency does NOT depend on n.

- Then $Eff(\tilde{(\theta)}, \hat{\theta})^{-1}$ tells you what sample size m will make $Var(\tilde{\theta}_m) \approx Var(\hat{\theta}_n)$.

13.3 Example: Tutorial 5, Q3

MOM estimator of θ, $\tilde{\theta} = \frac{7}{6} - \frac{\bar{X}}{2}$ is unbiased; it is not efficient.

- $Var(\tilde{\theta}) = \frac{1}{18n} + \frac{\theta(1-\theta)}{n}$
- $I(\theta) = \frac{1}{\theta(1-\theta)}$ — how did we get this?

- $Eff(\tilde{\theta}) = \frac{I(\theta)^{-1}/n}{Var(\tilde{(\theta)})} = (1 + \frac{1}{18\theta(1-\theta)})^{-1}$ — least efficient when $\theta = \frac{1}{2}$

ML estimator of θ, $\hat{\theta} = \frac{V_0+V_1}{n}$ is unbiased; it is also efficient.

- $Var(\hat{\theta}) = \frac{\theta(1-\theta)}{n} = \frac{I(\theta)^{-1}}{n}$

Relative efficiency

- $Eff(\tilde{\theta}, \hat{\theta}) = Eff(\tilde{\theta}) = (1 + \frac{1}{18\theta(1-\theta)})^{-1}$

Let sample size for ML be n, sample size for MOM be kn.

- $Var(\hat{\theta}) = \frac{\theta(1-\theta)}{n}$, $Var(\tilde{\theta}) = \frac{1}{18kn} + \frac{\theta(1-\theta)}{kn}$
- Equating the variances gives $k = 1 + \frac{1}{18\theta(1-\theta)}$, which is $Eff(\tilde{\theta}, \hat{\theta})^{-1}$
- For $\theta = 0.5$, we get $k \approx 1.22$, so the MOM needs 22% more samples than ML to have the same variance

13.4 Efficiency of *Consistent* Estimators (Asymptotic Efficiency)

Let $\tilde{\theta}$ and $\hat{\theta}$ be consistent estimators.

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!

Efficiency of ML estimators

- ($\star$) $Eff(\hat{\theta}) \to 1$ as $n \to \infty$!

13.5 Bias vs. Variance

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.

13.6 Choice of Estimator

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

- One way is to choose the one the lowest *MSE*: $MSE = E(\hat{\theta} - \theta)^2 = Var(\hat{\theta}) + (E(\hat{\theta}) - \theta)^2 = SE^2 + bias^2$
- Since MSE is 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

14 Sufficiency

(Recall: $f_{X|Y=y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$)

Let $X_1, \ldots, X_n$ be IID with density $f_\theta(x)$, $\theta \in \Theta$.

Let $T(\boldsymbol{X})$ be a function of $\boldsymbol{X} = (X_1, \ldots, X_n)$. In general, the conditional distribution of $\boldsymbol{X}$ given $T = t$ depends on θ.

14.1 Definition of 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 $T = t$

Intuitively, information about θ in $\boldsymbol{X}$ is all contained in T if it is sufficient.

14.2 Characterisation of Sufficiency

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 $\boldsymbol{x}$)

- For each t, let $S_t = \{\boldsymbol{x} \mid T(\boldsymbol{x}) = t\}$.
- The sample space of $\boldsymbol{X}$, S, is the disjoint union of S_t across all possible values of T.
- If $\boldsymbol{x} \notin S_t$, then $f_\theta(\boldsymbol{X} = \boldsymbol{x} \mid T = t) = 0$, which is not interesting

i.e. we reduce it to only look at $\boldsymbol{x} \in S_t$

Example: Sum of Bernoullis

Let $X_1, \ldots, X_n$ be IID $Ber(p)$, and let $T = X_1 + \ldots + X_n$.

For $t \in \{0, \ldots, n\}$, $S_t = \{\boldsymbol{x} \mid x_1 + \ldots + x_n = t\}$.

For $\boldsymbol{x} \in S_t$,

$$\begin{aligned} P_p(\boldsymbol{X} = \boldsymbol{x} \mid T = t) &= \frac{P_p(\boldsymbol{X} = \boldsymbol{x}, T = t)}{P_p(T = t)} = \frac{P_p(\boldsymbol{X} = \boldsymbol{x})}{P_p(T = t)} \\ &= \frac{P_p(\boldsymbol{X} = \text{a sequence of } t \text{ 1's}, n - t \text{ 0's})}{P_p(T = t)} \\ &= \frac{p^t(1-p)^{n-t}}{\binom{n}{t} p^t (1-p)^{n-t}} = \binom{n}{t}^{-1} =: q(\boldsymbol{x}) \end{aligned}$$

The conditional distribution of $\boldsymbol{X}$ given T is the same for every $p \in (0, 1)$, and this holds for every t. So T is sufficient for p.

Example of Insufficiency

Let X_1, X_2 be IID $Ber(p)$, and let $T = X_1$. $S_0 = \{(0,0), (0,1)\}$, $S_1 = \{(1,0), (1,1)\}$.

For $\boldsymbol{x} \in S_0$,

$$\begin{aligned} P_p(\boldsymbol{X} = \boldsymbol{x} \mid T = 0) &= \frac{P_p(X_1 = 0, X_2 = x_2)}{P_p(X_1 = 0)} \\ &= P_p(X_2 = x_2) \\ &= p^{x_2}(1-p)^{1-x_2} \end{aligned}$$

The conditional distribution of $\boldsymbol{X}$ given $T = 0$ depends on p. So T is not sufficient for p.

14.3 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 $\boldsymbol{x}$

Idea: 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

Proof

Proof that factorisation implies sufficiency:

Suppose that for every $\theta \in \Theta$, $P_\theta(\boldsymbol{X} = \boldsymbol{x}) = g(T(\boldsymbol{x}), \theta) \cdot h(\boldsymbol{x})$ for all possible $\boldsymbol{x}$.

We show that the characterisation holds. For all $\boldsymbol{x} \in S_t$:

$$\begin{aligned} P_\theta(\boldsymbol{X} = \boldsymbol{x} | T = t) &= \frac{P_\theta(\boldsymbol{X} = \boldsymbol{x}, T = t)}{P_\theta(T = t)} = \frac{P_\theta(\boldsymbol{X} = \boldsymbol{x})}{P_\theta(T = t)} = \frac{P_\theta(\boldsymbol{X} = \boldsymbol{x})}{\sum_{\boldsymbol{x}^* \in S_t} P_\theta(\boldsymbol{X} = \boldsymbol{x}^*)} \\ &= \frac{g(t, \theta) \cdot h(\boldsymbol{x})}{\sum_{\boldsymbol{x}^* \in S_t} g(t, \theta) \cdot h(\boldsymbol{x})} \\ &= \frac{h(\boldsymbol{x})}{\sum_{\boldsymbol{x}^* \in S_t} h(\boldsymbol{x}^*)} =: q(\boldsymbol{x}) \end{aligned}$$

This conditional distribution is the same for every $\theta \in \Theta$, and holds for every t, so T is sufficient for θ.

Proof that sufficiency implies factorisation:

Suppose that T is sufficient for θ, then it can be characterised: there is $q(\boldsymbol{x})$ such that for every $\theta \in \Theta$ and t, $P_\theta(\boldsymbol{X} = \boldsymbol{x} | T = t) = q(\boldsymbol{x})$, $\boldsymbol{x} \in S_t$.

We show the factorisation in each S_t. For $\boldsymbol{x} \in S_t$:

$$\begin{aligned} P_\theta(\boldsymbol{X} = \boldsymbol{x}) &= P_\theta(\boldsymbol{X} = \boldsymbol{x}, T = t) \\ &= P_\theta(T = t) P_\theta(\boldsymbol{X} = \boldsymbol{x} | T = t) \\ &= P_\theta(T = t) \cdot q(\boldsymbol{x}) \end{aligned}$$

Then let $g(t, \theta) = P_\theta(T = t)$, and $h(\boldsymbol{x}) = q(\boldsymbol{x})$. Hence $P_\theta(\boldsymbol{X} = \boldsymbol{x}) = g(T(\boldsymbol{x}), \theta) \cdot h(\boldsymbol{x})$.

Example: Bernoulli

Let $\boldsymbol{x}$ be realisations from IID $Bernoulli(p)$ random variables $X_1, \ldots, X_n$

$f_p(\boldsymbol{x}) = \sum_{i=1}^n p^{x_i}(1-p)^{1-x_i} = p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}$

Let $T = X_1 + \ldots + X_n$ and $t = \sum_{i=1}^n x_i$, then let $g(t, p) = p^t(1-p)^{n-t}$, $h(\boldsymbol{x}) \equiv 1$

Then for every p and $\boldsymbol{x}$, $f_p(\boldsymbol{x}) = g(T(\boldsymbol{x}), p) \cdot h(\boldsymbol{x})$

Example: Poisson

$$\begin{aligned} P_\lambda(\boldsymbol{X}=\boldsymbol{x}) &= \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} \\ &= \frac{e^{-n\lambda}\lambda^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n (x_i!)} \\ &= (e^{-n\lambda}\lambda^{\sum_{i=1}^n x_i}) \cdot \frac{1}{\prod_{i=1}^n (x_i!)} \end{aligned}$$

Hence $T(\boldsymbol{X}) = \sum_{i=1}^n X_i$ is sufficient for λ.

(Similarly, we can also show that $T(\boldsymbol{X}) = \bar{X}$ is sufficient.)

Example: Normal

$$\begin{aligned} f_{(\mu,\sigma^2)}(\boldsymbol{x}) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \\ &= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu)^2} \\ &= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n e^{-\frac{1}{2\sigma^2}[\sum_i x_i^2 - 2\mu\sum_i x_i + n\mu^2]} \end{aligned}$$

This shows that $(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)$ is sufficient for (μ, σ^2).

- Let $g(t,\theta)$ be everything as above, let $h(\boldsymbol{x}) \equiv 1$.
- Notice that here we can't really separate functions of $\boldsymbol{x}$ from our parameters μ and σ^2;

so the functions of $\boldsymbol{x}$ we need to keep here are $\sum_i x^2$ and $\sum_i x_i^2$

$$\ldots = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n e^{\frac{\mu}{\sigma}\sum_i x_i - \frac{n\mu^2}{2\sigma^2}} e^{-\frac{1}{2\sigma^2}\sum_i x_i^2}$$

This shows that if σ^2 is known, then $\sum_{i=1}^n X_i$ is sufficient for μ.

- Let $g(t,\theta)$ be the LHS, let $h(\boldsymbol{x})$ be $e^{-\frac{1}{2\sigma^2}\sum_i x_i^2}$.
- Because we can separate the $\sum_i x_i$ and μ on the LHS from the $\sum_i x_i^2$ on the RHS

14.4 Significance of Sufficiency

Theorem: If T is sufficient of θ, then the ML estimator is a function of T.

Proof: Given realisations $\boldsymbol{x}$, with summary t, the likelihood function is $g(t,\theta)h(\boldsymbol{x})$. (WHY?) The ML estimate is the θ value that maximises $g(t,\theta)$, so it is a function of t.

($\star$) Corollary: For an estimator $\hat{\theta}$ which is not a function of a sufficient statistic T, there is a better estimator by conditioning on T.

Example: Bernoulli

Let $X_1, \ldots, X_n$ be IID $Bernoulli(p)$. Let $T = X_1 + \ldots + X_n$ be sufficient for p.

Let $\hat{p} = X_1$ — this is not a very good estimator, but at least it's unbiased.

What is the conditional distribution of X_1 given $T = t \in \{0, 1, \ldots, n\}$?

$$\begin{aligned} P_p(X_1 = 1|T = t) &= \frac{P_p(X_1 = 1, X_2 + \ldots + X_n = t - 1)}{P_p(T = t)} \\ &= \frac{p \cdot \binom{n-1}{t-1} p^{t-1}(1-p)^{n-t}}{\binom{n}{t} p^t (1-p)^{n-t}} \\ &= \frac{t}{n} \\ P_p(X_1 = 0|T = t) &= 1 - \frac{t}{n} \end{aligned}$$

Then $(X_1|T = t) \sim Bernoulli(\frac{t}{n})$.

- $E(X_1|T = t) = \frac{t}{n}$
- $Var(X_1|T = t) = \frac{t}{n}(1 - \frac{t}{n})$

Now take the *random conditional expectation*: $E(X_1|T) = \frac{T}{n}$

- (We know distribution of $T \sim Bin(n, p)$)
- Using the old estimator, take at its condition expectation over sufficient statistic $T \rightarrow$ new estimator!
- $E(E(X_1|T)) = \frac{np}{n} = p$
- $Var(E(X_1|T)) = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}$
- $E(Var(X_1|T)) = p(1-p)(1 - \frac{1}{n}) = \frac{p(1-p)(n-1)}{n}$

14.5 Random Conditional Expectation

Idea: take the random conditional expectation of our old estimator over sufficient statistic T to give a new estimator

- $E(X_1|T) = \frac{T}{n}$
- $E(X_1) = E(E(X_1|T))$
- $Var(X_1) = Var(E(X_1|T)) + E(Var(X_1|T))$
- For Bernoulli example, $p(1-p) = \frac{p(1-p)}{n} + \frac{p(1-p)(n-1)}{n}$

14.6 General Definitions

- $E(Y|x)$ is the *conditional expectation* of Y given $X = x$
- $E(Y|X)$ is the *random conditional expectation* Y given X, which takes the value $E(Y|x)$ given $X = x$
- $Var(Y|x) = E(Y^2|x) - E(Y|x)^2$ is the *conditional variance* of Y given $X = x$
- $Var(Y|X)$ is the (random?) *conditional variance* of Y given X

Note that $E(Y) = Y$ and $Var(Y) = 0$ if and only if Y is a constant.

Similarly, $E(Y|X) = Y$ and $Var(Y|X) = 0$ if and only if Y is a function of X. Proof:

- Suppose $Y = g(X)$. Then for any $X = x$, $E(Y|x) = g(x)$, so $E(Y|X) = g(X) = Y$ and $Var(Y|X) = 0$
- Suppose $E(Y|X) = Y$. Then for any $X = x$, $E(Y|x) = Y$, so Y is a constant when $X = x$, i.e. Y is a function of X.

Then for any x, $Var(Y|x) = 0$, so Y is a constant when $X = x$.

Let $Y = g(X)$.

- $E[E(g(X)|X)] = g(X)$
- $Var[E(g(X)|X)] + E[Var(g(X)|X)] = Var(g(X))$

Two more important facts about conditional expectations:

1. $E(Y) = E[E(Y|X)]$
   - See proof in Rice p149
2. $Var(Y) = Var[E(Y|X)] + E[Var(Y|X)]$
   - Let $Z_k = E(Y^k|X)$. $Var(Y|X) = Z_2 - Z_1^2$
   - $Var[E(Y|X)] = Var(Z_1) = E(Z_1^2) - E(Z_1)^2 = E(Z_1^2) - E(Y)^2$ (by 1)
   - $E[Var(Y|X)] = E(Z_2 - Z_1^2) = E(Z_2) - E(Z_1^2) = E(Y^2) - E(Z_1^2)$ (by 1)

Example

Let $X \sim N(\mu, \sigma^2)$ and $\epsilon \sim N(0, \tau^2)$ be independent, let $Y = X + \epsilon \sim N(\mu, \sigma^2 + \tau^2)$. Verify the previous facts:

1. $E(Y|X) = E(X + \epsilon|X) = X + E(\epsilon|X)$. Then $E[E(Y|X)] = E(X) = \mu$.
2. $Var(Y|X) = Var(X + \epsilon|X) = 0 + Var(\epsilon|X) = \tau^2$. Then $Var[E(Y|X)] + E[Var(Y|X)] = Var(X) + E(\tau^2) = \sigma^2 + \tau^2 = Var(Y)$.

14.7 Rao-Blackwell Theorem

Idea: your new estimator, made by taking the conditional expectation on T, will be strictly better than your old estimator in terms of MSE.

Let $\hat{\theta}$ be an estimator of θ with finite variance, let T be sufficient for θ. Define $\tilde{\theta} = E(\hat{\theta}|T)$. Then for every $\theta \in \Theta$,

$$E(\tilde{\theta} - \theta)^2 \leq E(\hat{\theta} - \theta)^2$$

(smaller MSE!)

(That $\tilde{\theta}$ is a function of T complements the fact that ML estimator is a function of T)

Proof:

- $E(\tilde{\theta}) = E[E(\hat{\theta}|T)] = E(\hat{\theta})$, so estimators have the same bias
- $Var(\hat{\theta}) = Var[E(\hat{\theta}|T)] + E[Var(\hat{\theta}|T)] = Var(\tilde{\theta}) + E[Var(\hat{\theta}|T)] \geq Var(\tilde{\theta})$
- It is only the case that $Var(\hat{\theta}) = Var(\tilde{\theta})$ when $E[Var(\hat{\theta}|T)] = 0$, i.e. it's already based on a sufficient statistic T
- Note: we need T to be sufficient so that $\tilde{\theta} = E(\hat{\theta}|T)$ does not depend on the unknown θ, and so $\tilde{\theta}$ is a well-defined estimator

15 Hypothesis Testing

Let $X_1, \ldots, X_n$ be IID random variables with unknown mean μ, density $f(x|\theta)$ where $\theta \in \Theta$ is an unknown constant

Question: does the data support or refute our hypothesis that θ equals some value? To answer this, we use statistical tests

15.1 Definitions

Null hypothesis H_0: the default belief, e.g. $\mu = 40$.

- ($\star$) In a hypothesis test, our goal is decide whether to either *reject* or *NOT reject* the null hypothesis.

Alternative hypothesis H_1: the other belief, e.g. $\mu = 43$.

The statistical question is to choose between the null and alternative hypothesis.

Critical region: a region of sample space, whereby if the data falls in the critical region, we *reject* the null hypothesis. E.g. $\{x > c\}$

Type I error: rejecting H_0 when it is true.

- *Size* i.e. *significance level*, is size $= P_0(X > c)$

Type II error: not rejecting H_0 when it is false.

- $P(\text{type II error}) = P_1(X < c)$
- power $= P_1(X > c) = 1 - P_1(\text{type II error})$

Size and power are of the same "form"; both deal with the critical region in rejecting the null hypothesis

- You want size to be small: size is probability of rejection under *null*
- You want power to be big: power is probability of rejection under *alternative*
- E.g. size is $P_0(X > c)$, power is $P_1(X > c)$

How do we choose an optimal critical region value?

- Trade-off between size and power

Control Size, Maximise Power

Neyman-Pearson approach: among tests with *size* less than α, choose the *most powerful* test.

- We're more concerned with making sure that α is small, since H_0 is the "default" belief; given that H_0 is true, we then

go about finding the one with the most power.

- If H_0 is true, we have a small chance α of rejecting it.
- If H_1 is true, we have as high a chance of rejecting H_0 as possible.

15.2 General Setup

Let $X_1, \ldots, X_n$ be IID with density $f(x|\theta)$. Suppose:

- $H_0 : \theta = \theta_0$
- $H_1 : \theta = \theta_1$

Let the critical region of a test be $R \subset \mathbb{R}^n$. Then:

- Size $= P_0(\mathbb{X} \in R)$
- Power $= P_1(\mathbb{X} \in R)$

Generally, it is hard to see if there is a most powerful test among those with size $\leq \alpha$.

Example: Amount of Custard in Egg Tart

- For the *baker*, the amount of custard in each egg tart he makes is $X \sim N(40, 2^2)$
- For the *trainee*, the amount of custard in each egg tart he maes is $N(43, 2^2)$
- Question: given the amount of custard from a tart, can we tell if the baker or trainee made it?

Setup

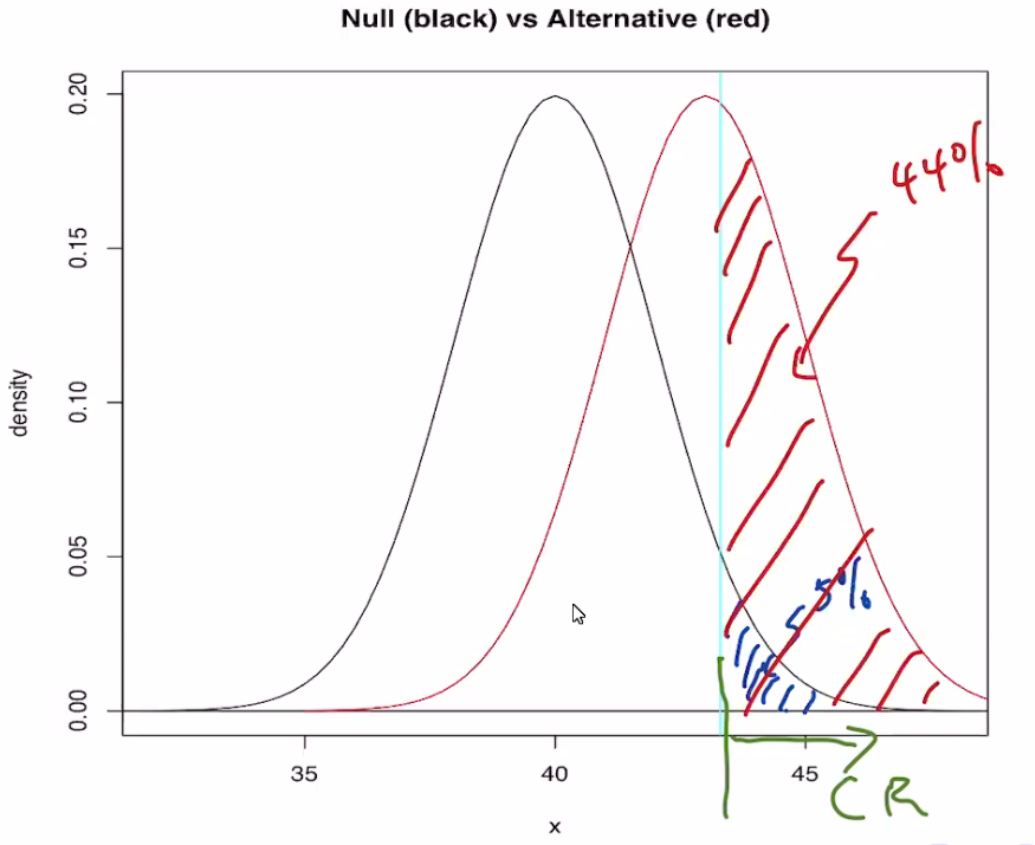
- Null hypothesis H_0: $\mu = 40$ (the baker made it)
- Alternative hypothesis H_1: $\mu = 43$ (the baker made it)

Let's have the critical region be $\{x > c\}$. Suppose $c = 42$.

- size $= P_0(X > 42) = P(Z > 1) \approx 0.16$
- power $= P_1(X > 42) = P(Z > -0.5) \approx 0.69$

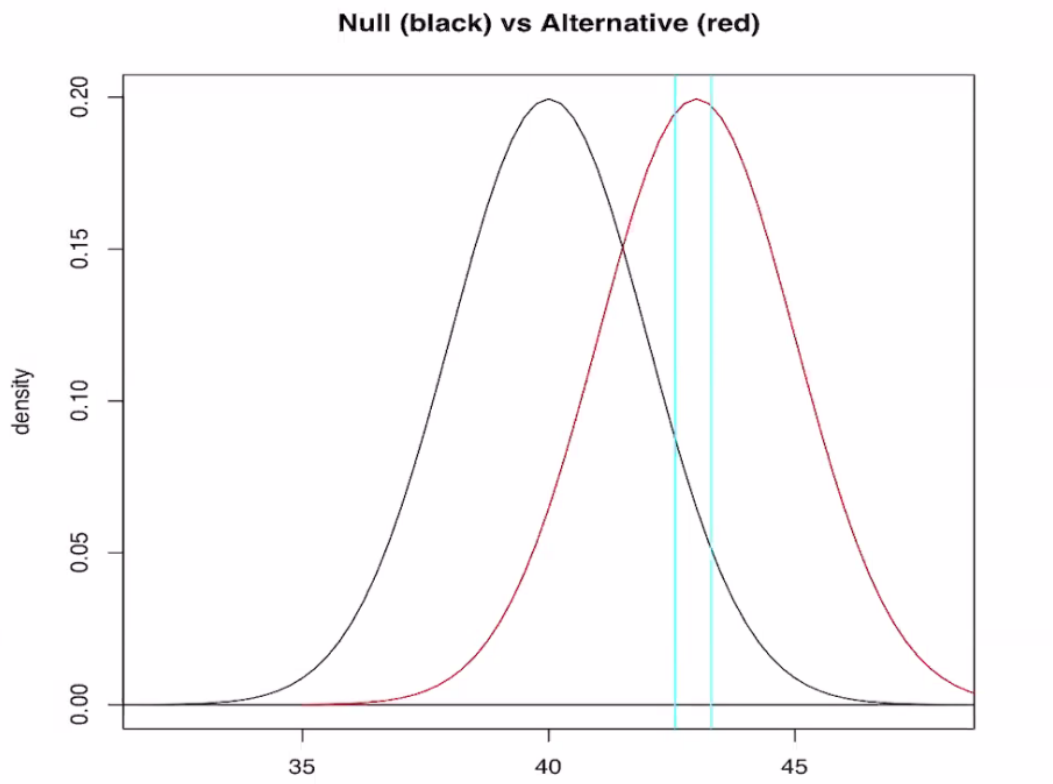
Suppose we take the test i.e. critical region to be $\{x > 43.3\}$.

- Then size $= 0.05$, power $= 0.44$; shown in the figure below



Suppose we take the test i.e. critical region to be $\{42.6 < x < 43.3\}$.

- Note that size is stil $= 0.05$, but power $= 0.14$ (worse!); shown in the figure below



Example: Normal

$X_1, \ldots, X_n$ are IID $N(\mu, \sigma^2)$ with known σ^2 and unknown μ.

Let $H_0 : \mu = \mu_0$. Possibilities for H_1:

- $\mu = \mu_1$, where $\mu_1 \neq \mu_0$
- $\mu > \mu_0$
- $\mu \neq \mu_0$

Consider $n = 1$, i.e. only one random variable.

- $H_0 : \mu = \mu_0$, $H_1 = \mu = \mu_1$, where $\mu_0 < \mu_1$

$$\wedge(x) = \frac{\exp\{-\frac{(x-\mu_0)^2}{2\sigma^2}\}}{\exp\{-\frac{(x-\mu_1)^2}{2\sigma^2}\}} = \exp\{-\frac{2x(\mu_1-\mu_0)-(\mu_1^2-\mu_0^2)}{2\sigma^2}\}$$

15.3 Likelihood Ratio

Likelihood ratio of H_0 to H_1 based on $X_1, \ldots, X_n$ is

$$\Lambda(\boldsymbol{x}) = \frac{f_0(x_1)\ldots f_0(x_n)}{f_1(x_1)\ldots f_1(x_n)}$$

The smaller $\Lambda(\boldsymbol{x})$ is, the more evidence there is against H_0 (or for H_1). A reasonable critical region consists of $\boldsymbol{x}$ with small $\Lambda(\boldsymbol{x})$.

15.4 Likelihood ratio tests

Consider critical regions to be $R_c = \{\boldsymbol{x} \mid \Lambda(\boldsymbol{x}) < c\}$, where $c > 0$.

As c increases, then what happens to size and power? R_c gets bigger, so size and power also increases.

For any $\alpha \in (0, 1)$, there is a c_α such that size is α.

15.5 Neyman-Pearson Lemma

Consider the likelihood ratio test with critical region $\{\boldsymbol{x} \mid \Lambda(\boldsymbol{x}) < c_\alpha\}$.

Among all tests with size $\leq \alpha$, this test has the maximum power.

Normal Case (A) n Normals, $\mu_1 > \mu_0$

Assume $\mu_0 < \mu_1$ and σ^2 is known.

$\wedge(\boldsymbol{x}) = \frac{\exp\{-\frac{\sum_{i=1}^n (x_i-\mu_0)^2}{2\sigma^2}\}}{\exp\{-\frac{\sum_{i=1}^n (x_i-\mu_1)^2}{2\sigma^2}\}} = \exp\{-\frac{2n\bar{x}(\mu_1-\mu_0)-n(\mu_1^2-\mu_0^2)}{2\sigma^2}\}$

By Neyman-Pearson lemma, the most powerful test of size $\leq \alpha$ has critical region of the form $\{\boldsymbol{x} \mid \Lambda(\boldsymbol{x}) < c_\alpha\}$.

In the normal case, this is the same as $\{\boldsymbol{x} \mid \bar{x} > c'_\alpha\}$ for some related constant c'_α. (Why? See tutorial question.)

What is c'_α?

$\alpha = P_0(\bar{X} > c'_\alpha) = P(Z > \frac{c'_\alpha - \mu_0}{\sigma/\sqrt{n}})$

Hence,

$\frac{c'_\alpha - 0}{\sigma/\sqrt{n}} = z_\alpha$, $c'_\alpha = \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}}$

What is the power?

$P_1(\bar{X} > c'_\alpha) = P(Z > \frac{c'_\alpha - \mu_1}{\sigma/\sqrt{n}}) = P(Z > z_\alpha - \frac{\mu_1 - \mu_0}{\sigma/\sqrt{n}})$

i.e. smaller $\sigma/\sqrt{n}$ i.e. larger n will increase our power!

Simple vs Composite Hypotheses

Simple: hypothesis completely specifies the distribution of the data

Composite: hypothesis DOES NOT completely specify the distribution of the data

- $\mu > \mu_0$ or $\mu < \mu_0$ or $\mu \neq \mu_0$

Normal Case (B) $\mu > \mu_0$ (one-sided)

Idea: reject H_0 if $\bar{x}$ is too large.

Critical region for size α: $\{\bar{x} > \mu_0 + z_\alpha \frac{\alpha}{\sqrt{n}}\}$

Power of the test depends on the value of $\mu > \mu_0$; so the power is now a *function* of μ

By Neyman-Pearson lemma, for any alternative $\mu_1 > \mu_0$, this test is the most powerful as its power *function* is the largest.

Uniformly most powerful test, i.e. for any other test of size $\leq \alpha$, its power function is smaller everywhere.

Normal Case (C) $\mu \neq \mu_0$ (two-sided)

How do we take into account not only $\mu > \mu_0$, but also $\mu < \mu_0$?

Idea: reject H_0 if $\bar{x}$ is too far away from μ_0 in both sides.

Critical region for size α: $\{|\bar{x} - \mu_0| < c\}$

$c = z_{\alpha/2} \frac{\alpha}{\sqrt{n}}$

This test is NOT uniformly most powerful: by Neyman-Pearson lemma, for any alternative if $\mu_1 > \mu_0$, there is a more powerful test.

Summary of Neyman-Pearson Lemma

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.

Likelihood ratio test is still very useful in many problems where the most powerful test may be powerful to get.

Example: Custard Tarts

Let $n = 25$, $\alpha = 0.05$ so $z_\alpha = 1.64$. Let $H_0 : \mu = 40$ and $H_1 : \mu = 43$.

Most powerful test of size $\leq \alpha$ has critical region $\bar{x} > 40 + 1.64 \times \frac{2}{\sqrt{25}} = 40.7$

Suppose that $\bar{x} = 40.3$, then we do not reject H_0 at level α.

95%-CI for μ: $40.3 \pm 1.96 \times \frac{2}{\sqrt{25}} = (39.5, 41.1)$

Duality of hypothesis tests and confidence intervals

$(1 - \alpha)$ CI for μ contains precisely the values μ_0 for which $H_0 : \mu = \mu_0$ is not rejected by two-sided test of size α, against $H_1 : \mu \neq \mu_0$.

i.e. complement set of critical region in two-sided test

15.6 *p*-value

Previously, we fixed an α at say 0.05 or 0.01, which then decides the critical region and hence whether to reject H_0 or not.

But the choice of α is arbitrary. Are there any other options?

Another way is to calculate p-value: the probability under H_0 that the test statistic (e.g.) is more extreme than the realisation.

E.g. bakery example: $p = P_0(\bar{X} > 40.3) = P(Z > \frac{40.3-40}{2/5}) \approx 0.23$

The smaller the p-value, the more likely one is to intuitively reject H_0

Note, p-value is NOT the probability that H_0 is true; either H_0 is true or H_0 is false, there is no probability distribution. What we're computing is that *assuming* H_0 is true, how likely is it that we get our θ.

***p*-value and size of test α**

If size of test is smaller than p-value, then H_0 will not be rejected. If size of test is larger than p-value, then H_0 will be rejected.

So p-value is the smallest α such that test of size α will be rejected. (E.g. if $p = 0.2$, then any $\alpha > 0.2$ size test will not reject H_0.)

Example: Normal Case

(A) and (B): $p = P_0(\bar{X} > \bar{x}) = P(Z > \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}})$

(C): $p = P_0(|\bar{X} - \mu_0| > |\bar{x} - \mu_0|)$

16 Generalised Likelihood Ratio Test

Let H_0 and H_1 be represented by subsets ω_0 and ω_1 of the parameter space Θ.

- E.g. Normal case (C): $X_1, \ldots, X_n$ IID $N(\mu, \sigma^2)$ with known σ^2. $H_0 : \mu = \mu_0$, $H_1 : \mu \neq \mu_0$. So $\omega_0 = \{\mu_0\}$, $\omega_1 = \{\mu | \mu \neq \mu_0\}$
- E.g. $\boldsymbol{X} \sim Multinomial(n, \boldsymbol{p})$. $H_0 : \boldsymbol{p} \in \omega_0$, $H_1 : \boldsymbol{p} \in \omega_1$, where ω_0 and ω_1 are disjoint sets of probability vectors of length r
- E.g. Poisson: $X_1, \ldots, X_n$ independent, where $X_i \sim Po(\lambda_i)$. $H_0 : \lambda$ are the same, $H_1 : \lambda$ are different.

So $\omega_0 = \{(\lambda_1, \ldots, \lambda_n) \mid \lambda_1 = \ldots = \lambda_n\}$, $\omega_1 = \{(\lambda_1, \ldots, \lambda_n) \mid \lambda_i \neq \lambda_j \text{ for some } i, j \text{ in } 1 \ldots n\}$

Generalised Likelihood Ratio

Let $L(\theta)$ be the likelihood function based on the data.

One way: $\Lambda^* = \frac{\max_{\theta \in \omega_0} L(\theta)}{\max_{\theta \in w_1} L(\theta)}$

Neyman-Pearson idea: small Λ^* value is evidence for H_1.

Let $\Omega = \omega_0 \cup \omega_1$; it is more convenient to use

$\Lambda = \frac{\max_{\theta \in \omega_0} L(\theta)}{\max_{\theta \in \Omega} L(\theta)}$

Note that $\Lambda \leq \Lambda^*$ and $0 \leq \Lambda \leq 1$. The closer Λ is to 0, the stronger the evidence is for H_1.

16.1 Large-sample null distribution of Λ

($\star$) Under H_0, if n is large, approximately $-2 \log \Lambda \sim \chi^2_k$, where $k = dim(\Omega) - dim(\omega_0)$ (dimension refers to number of ways we can change the parameters freely)

- E.g. $Multinomial(n, p)$ has dimension of $n - 1$ (last one is wholly determined by the previous $n - 1$ ones)

Example: Normal Case (C)

$\omega_0 = \{\mu_0\}$, $\omega_1 = \{\mu \mid \mu \neq \mu_0\}$, so $\Omega = \mathbb{R}$

$$\max_{\mu \in \omega_0} L(\mu) = \frac{1}{\sigma^n (2\pi)^{n/2}} \exp\{-\frac{\sum_{i=1}^n (X_i - \mu_0)^2}{2\sigma^2}\}$$

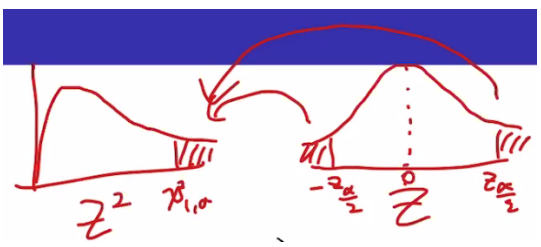
$$\max_{\mu \in \Omega} L(\mu) = \frac{1}{\sigma^n (2\pi)^{n/2}} \exp\{-\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{2\sigma^2}\} \text{ (recall } \sum_i (x_i - c) \text{ is minimized at } c = \bar{x})$$

$$-2 \log \Lambda = \ldots = \frac{n(\bar{X} - \mu_0)^2}{\sigma^2} = \left(\frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}\right)^2$$

Under H_0, $\bar{X} \sim N(\mu_0, \sigma^2/n)$, so $-2 \log \Lambda \sim \chi^2_1$ for any n. Check $k = 1$.

We want to reject for small Λ, i.e. reject for large $-2 \log \Lambda$.

For size α, critical region is $\{\frac{(\bar{x} - \mu_0)^2}{\sigma^2/n} > \chi^2_{1,\alpha}\} = \{|\bar{x} - \mu_0| > z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\}$



Let the realised sample mean be $\bar{x}$. The p-value is $P(\chi_1^2 > \frac{(\bar{x}-\mu_0)^2}{\sigma^2/n})$

Example: Multinomial

Let ω_0 consist of $\boldsymbol{p}(\theta)$, where $\theta \in \Theta$ is a constant. Let ω_1 consist of all $\boldsymbol{p}$ not in ω_0, so Ω is the set of all probability vectors.

Let $\hat{\theta}$ be the ML estimator of θ. Then ML estimator of $\boldsymbol{p}$ under ω_0 is $\boldsymbol{p}(\hat{\theta})$; ML estimator of $\boldsymbol{p}$ under Ω is $\frac{\boldsymbol{X}}{n}$

$$\max_{\boldsymbol{p}\in\omega_0} L(\boldsymbol{p}) = \binom{n}{X_1 \dots X_r} \prod_{i=1}^{r} p_i(\hat{\theta})^{X_i}$$

$$\max_{\boldsymbol{p}\in\Omega} L(\boldsymbol{p}) = \binom{n}{X_1 \dots X_r} \prod_{i=1}^{r} p_i(\frac{X_i}{n})^{X_i}$$

$$\Lambda = \prod_{i=1}^{r} \left(\frac{np_i(\hat{\theta})}{X_i}\right)^{X_i} = \prod_{i=1}^{r} \left(\frac{E_i}{X_i}\right)^{X_i}$$

$$-2\log\Lambda = 2\sum_{i=1}^{r} X_i \log\left(\frac{X_i}{E_i}\right)$$

$$\approx \sum_{i=1}^{r} \frac{(X_i - E_i)^2}{E_i} \text{ (for large } n)$$

the Pearson's chi-squared statistic X^2.

where $E_i = np_i(\hat{\theta})$ is the expected frequency of the i-th event under H_0.

Example: Multinomial HWE

$\boldsymbol{p} \in \omega_0 = \{((1-\theta)^2, 2\theta(1-\theta), \theta^2) \mid \theta \in (0,1)\}$ — dimension 1

Let Ω be the set of all probability vectors, of dimension 2

Recall our observed data with $n = 1029$:

- $x_1 = 342$, $x_2 = 500$, $x_3 = 187$

Since ML estimate of θ is 0.4247, expected counts are:

- Of AA: $E_1 = 1029 \times 0.5753^2 \approx 340.6$
- Of Aa: $E_2 \approx 502.8$
- Of aa: $E_3 \approx 185.6$

So $-2\log\Lambda \approx X^2 \approx 0.0319$.

$P(\chi_1^2 > 3.84) \approx 0.05$ => critical region is > 3.84. Since $n = 1029$ is large, we do not reject H_0 at $\alpha = 0.05$.

p-value is $P(\chi_1^2 > 0.0319) \approx 0.86$.

Example: Rice page 344 Example B

Assume a Poisson model

- $H_0 : \boldsymbol{x}$ is determined by Poisson distribution — ω_0 has dimension 1
- $H_1 : \boldsymbol{x}$ is arbitrary — Ω has dimension 7 (last one, the eighth probability, is determined by previous seven)

We get the estimate for λ to be 2.44.

Chi-square statistic $X^2 = 75.4$.

First, in our data, we group the tail categories together to give 8 remaining categories, otherwise n is too small (Rule of thumb: expected counts should be at least 5 for good approximation?)

Since $P(\chi_6^2 > 18.55) \approx 0.05$, our p-value is less than 0.05, so we reject the Poisson model.

Example: Poisson dispersion test

Let $X_i \sim Po(\lambda_i)$ be independent.

- H_0 : the λ are the same
- H_1 : the λ are different
- $\omega_0 = \{(\lambda_1, \ldots, \lambda_n) \mid \lambda_1 = \ldots = \lambda_n\}$
- $\omega_1 = \{(\lambda_1, \ldots, \lambda_n) \mid \lambda_i \neq \lambda_j\}$
- In ω_1, the MLE is $\hat{\lambda} = \bar{X}$
- In Ω, the MLEs are $\hat{\lambda}_i = X_i$

$$\begin{aligned}
\max_{\theta \in \omega_0} L(\theta) &= \prod_{i=1}^{n} \bar{X}^{X_i} e^{-X} / X_i! \\
\max_{\theta \in \Omega} L(\theta) &= \prod_{i=1}^{n} X^{X_i} e^{-X} / X_i! \\
\Lambda &= \prod_{i=1}^{n} \left(\frac{\bar{X}}{X_i} \right)^{X_i} e^{X_i - \bar{X}} \\
-2 \log \Lambda &= -2 \sum_{i=1}^{n} \left[X_i \log \left(\frac{\bar{X}}{X_i} \right) + X_i - \bar{X} \right] \\
&= 2 \sum_{i=1}^{n} X_i \log \left(\frac{X_i}{\bar{X}} \right)
\end{aligned}$$

It can be shown that $-2 \log \Lambda \approx \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{\bar{X}}$

Under H_0, numerator is around $(n-1)\lambda$, denominator is around λ (whut lol)

17 Comparing 2 Samples: Independent Samples

17.1 Normal Theory: Same Variance

$X_1, \ldots, X_n \sim N(\mu_X, \sigma^2)$ and $Y_1, \ldots, Y_n \sim N(\mu_Y, \sigma^2)$, independent

Test $H_0 : \mu_X = \mu_Y$ vs $H_1 : \mu_X \neq \mu_Y$

Test statistic: $\bar{X} - \bar{Y}$, which is normal

- $E(\bar{X} - \bar{Y}) = \mu_X - \mu_Y$
- $Var(\bar{X} - \bar{Y}) = \sigma^2 \left(\frac{1}{n} + \frac{1}{m}\right)$

If variance σ^2 is known

Define $Z = \frac{\bar{X} - \bar{Y}}{\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}}$

- Reject H_0 when $|Z| > z_{\alpha/2}$

More generally, to test for $H_0 : \mu_X - \mu_Y = d$, we use:

- $Z = \frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}}$

If variance σ^2 is unknown

Estimate σ^2 by the *pooled sample variance*:

- $s_p^2 = \frac{(n-1)s_X^2 + (m-1)s_Y^2}{m+n-2}$, where $s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$
- s_p^2 is an unbiased estimator of σ^2

Distribution of s_p^2: χ^2_{m+n-2}

t-statistic

Define $t = \frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{s_p\sqrt{\frac{1}{n} + \frac{1}{m}}}$: has a t distribution with $m + n - 2$ degrees of freedom

Confidence Intervals

Construct $(1 - \alpha)$ CI for $(\mu_X - \mu_Y)$

$(\bar{X} - \bar{Y}) \pm z_{\alpha/2} \cdot \sigma\sqrt{\frac{1}{n} + \frac{1}{m}}$ (if σ is known)

$(\bar{X} - \bar{Y}) \pm t_{m+n-2,\alpha/2} \cdot s_p\sqrt{\frac{1}{n} + \frac{1}{m}}$ (if σ is unknown)

One-sided vs Two-sided tests

Two-sided tests where $H_1 : \mu_X \neq \mu_Y$ was covered earlier: reject when $|t| > t_{n+m-2,\alpha/2}$ ($\alpha/2$ level)

One-sided tests where $H_1 : \mu_X > \mu_Y$: reject when $t > t_{n+m-2,\alpha}$ (α level)

Hence a two-sided test at α level is equivalent to one-sided test at $\alpha/2$ level. Important part is not to choose which test to use, but to clearly report which test you use.

17.2 Normal Theory: Unequal Variance

If variances 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 variances 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 degrees of freedom *df*

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

17.3 Summary

When variances are known, use Z-test

When variances are unknown, use t-test

- If we assume variances are the same, estimate σ^2 with s_p^2
- If we assume variances are different, estimate σ_X^2 and σ_Y^2 with s_X^2 and s_Y^2
- We may assume variances are the same if we check that s_X^2 and s_Y^2 are not too different; i.e. s_X and s_Y are within factor of 2

What if samples are NOT drawn from normal distributions?

- $\bar{X} - \bar{Y}$ still approximately normal by CLT if n and m are large
- Also, if n and m are large, t-statistics are also approximately normal,

so even in the unknown variance case we sometimes call it the Z-test

Example: Mining

Mine 1: 8260, 8130, 8350, 8070, 8340 Mine 2: 7950, 7890, 7900, 8140, 7920, 7840

Use 1% LOS to test whether the means are different.

- $H_0 : \mu_1 = \mu_2$
- $H_1 : \mu_1 \neq \mu_2$

Small sample size, so *assume* they are from a normal distribution.

Data:

- $s_1 \approx 125.5$
- $s_2 \approx 104.5$
- Since $\frac{1}{2} < s_1/s_2 < 2$, we use equal variance assumption
- $\bar{x} = \ldots$, $\bar{y} = \ldots$

T-test:

- $t = \frac{\bar{X} - \bar{Y} - 0}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$

- Rejection region: $|t| > t_{n_1+n_2-2,0.01/2} = 3.250$
- Plugging in everything, we get $t = 4.19$
- Since $t = 4.19$ is in the critical region, we reject H_0

17.4 Non-Parametric Test

Now we do not assume that X_i and Y_i are normal.

Let $X_1, \ldots, X_n$ be IID with CDF F, let $Y_1, \ldots, Y_n$ be IID with CDF G, X and Y are independent

Consider the null hypothesis $H_0 : F = G$.

We are interested in whether X values are larger than Y values on the whole, or vice versa: often phrased as $H_0 : \mu_X = \mu_Y$. Note that when we reject the former, we don't necessarily reject the latter, but in most practical situations we think of them as equivalent.

17.5 Mann-Whitney (Wilcoxon) Test

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, we can assign them the *average* of the ranks; if there are only a few tied values, significance levels not greatly affected
- For large m and n, why not just use Z-test? If you think the normal approximation is not good enough (m and n not large enough for a very bad underlying distribution)

Define $Rank(Z) = i$ if Z is the i-th *smallest* value within the pooled sample.

Define the rank sum scores:

- $T_X = \sum_{i=1}^{n} Rank(X_i)$
- $T_Y = \sum_{i=1}^{m} Rank(Y_i)$
- Note that $T_X + T_Y = \sum_{i=1}^{m+n} i = \frac{(m+n)(m+n+1)}{2}$ is fixed

Idea: if $H_0 : F = G$ is true, the ranks should be uniformly distributed from $\{1, \ldots, m+n\}$, so the rank sums should not be too small or too large.

Take the smaller sample of size $n_1 = min(m, n)$, and compute sum of ranks R from that. Then $R' = n_1(m + n + 1) - R$.

Mann-Whitney test statistic: $R^* = min(R, R')$

- Reject $H_0 : F = G$ if R^* is too small
- Can be either one-sided or two-sided test

Example: Mining

Mine 1: ranks are 9, 7, 11, 6, 10 Mine 2: ranks are 5, 2, 3, 8, 4, 1

Then:

- $R = 9 + 7 + 11 + 6 + 10 = 43$
- $R' = 5(12) - 43 = 17$
- $R^* = \min(R, R') = 17$

Rejection region at $\alpha = 0.01$ is $R \leq 16$, so do not reject H_0. (actual p-value is 0.017)

Performance of Mann-Whitney Test

- Works for *all* distributions, not only normal (non-parametric)
- Robust against outliers (unlike normal tests)
  - So more powerful than t-test: this is because outliers in t-test inflate sample variance, resulting in small t-value
- If the underlying distribution is indeed normal, this test is only slightly less efficient than 2-sample t-test
  - Asymptotic efficiency is about 0.955 => to achieve the same power as 2-sample t-test, need only about 5% more samples

18 Comparing 2 Samples: Paired Samples

18.1 Normal Theory

In paired samples, X and Y values are *paired up*, related to the same individual/object

- Assume that (X_i, Y_i) is independent of (X_j, Y_j)
- Let $D_i = Y_i - X_i$
- Let μ_D be unknown population mean of D values
- To test $H_0 : \mu_D = 0$, compute t-test statistic $t = \frac{\bar{D}}{s_D/\sqrt{n}}$, 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)
- If n is large, don't need normal assumption and reject H_0 when $|t| > z_{\alpha/2}$

In general, to test $H_0 : \mu_D = d$, we use:

- $t = \frac{\bar{D}-d}{s_D/\sqrt{n}}$
- $(1-\alpha)$ CI for μ_D: $\bar{D} \pm t_{n-1,\alpha/2} s_D/\sqrt{n}$

Example: Smoking

ID	1	2	3	4	5	6
X	25	25	27	44	30	60
Y	27	31	37	56	43	57

2 6 10 12 13 -3 4 36 100 144 169 9 sum=40 sum^2=462 n=6 462-(1600/6)

Test the hypothesis that $H_0 : \mu_D = 0$ (i.e. $\mu_X = \mu_Y$)

- $\bar{D} = 6.67$
- $s_D = 6.25$
- $t = \frac{6.67}{6.25/\sqrt{6}} = 2.61$
- This value is in the rejection region $|t| > t_{5,0.025} = 2.57$, so reject H_0

18.2 Non-Parametric test: Wilcoxon Signed-Rank Test

Wilcoxon signed-rank test based on differences D_i

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

Performing the test

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

Example: Smoking

ID	1	2	3	4	5	6
$D = Y - X$	2	6	10	12	13	-3
$Rank(D)$	1	3	4	5	6	2

- From the data, $W_+ = 19$, and $W_- = 2$
- Rejection region for $\alpha = 0.05$ is $W \leq 0$ (we don't have enough data points!), hence we fail to reject H_0