1 Probability Review

Multinomial Distribution

$$\Pr(X_1 = x_1, \ldots, X_r = x_r) = \binom{n}{x_1, \ldots, x_r} \prod_{i=1}^{r} p_i^{x_i}$$

Mean Square Error (MSE)

$$\mathsf{E}\{(Y-c)^2\} = \mathsf{var}(Y) + \{\mathsf{E}(Y) - c\}^2$$

$$\mathsf{E}\{(Y-c)^2|x\} = \mathsf{var}[Y|x] + \{\mathsf{E}[Y|x] - c\}^2$$

which are special cases of $\mathsf{E}(Y^2) = \mathsf{var}(Y) + [\mathsf{E}(Y)]^2$. MSE is minimized if and only if $c = \mathsf{E}(Y)$ or $\mathsf{E}[Y|x]$.
Usually the formula for $\mathsf{E}[Y|x] = f(x)$ is determined from observations/data and x can be a vector of realisations from covariates.

$$\mathsf{MSE}_{\mathsf{empirical}} = \frac{1}{n}\sum_{i=1}^{n}\{\mathsf{E}[Y|x_i] - y_i\}^2$$

In the real world, we have different realisations x_i of the random variable X, hence the mean MSE is

$$\frac{1}{n}\sum_{i=1}^{n}\mathsf{var}[Y|x_i] \approx \mathsf{E}(\mathsf{var}[Y|X]) \leq \mathsf{var}(Y)$$

Analysis of Variance (ANOVA)
involves breaking of variance into components

$$\mathsf{var}(Y) = \mathsf{E}(\mathsf{var}[Y|X]) + \mathsf{var}(\mathsf{E}[Y|X])$$

1.1 Distributions

χ_1^2 distribution
Let $Z \sim \mathcal{N}(0,1)$. $V = Z^2$ has a χ^2 distribution with 1 degree of freedom

$$f(v) = \frac{1}{\sqrt{2\pi}} v^{-1/2} e^{-v/2}$$

Gamma distribution

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

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx$$

χ_n^2 distribution
Let $V_1, \ldots, V_n$ be IID χ_1^2

$$V = \sum_{i=1}^{n} V_i$$

has a χ_n^2 distribution with n degrees of freedom

t distribution
Let $Z \sim \mathcal{N}(0,1)$ and $V \sim \chi_n^2$ be independent

$$t_n = \frac{Z}{\sqrt{V/n}}$$

has a t distribution with n degrees of freedom

F distribution
Let $V \sim \chi_m^2$ and $W \sim \chi_n^2$ be independent

$$F_{m,n} = \frac{V/m}{W/n}$$

has an F distribution with (m, n) degrees of freedom
*Note: $t_n^2 = F_{1,n}$

1.2 Sample Variance

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

$\bar{X}$ and S^2 are independent

$$\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2$$

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

2 Survey and Random Sampling

Let $X_1, \ldots, X_N$ be random draws without replacement from a population of size N with mean μ and variance σ^2.

$$\mathsf{cov}(X_i, X_j) = -\frac{\sigma^2}{N-1} \forall i \neq j$$

$$\mathsf{var}(\bar{X}) = \left(\frac{N-n}{N-1}\right)\frac{\sigma^2}{n}$$

2.1 Exchangeable

RV's $Y_1, \ldots, Y_k$ are exchangeable if all reordered vectors have the same distribution as $(Y_1, \ldots Y_k)$. i.e. for any permutation π on $\{1, \ldots, K\}$,

$$(Y_{\pi(1)}, \ldots, Y_{\pi(k)}) \stackrel{d}{=} (Y_1, \ldots, Y_k)$$

2.2 Estimate and Estimator

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- μ, σ, σ^2 are **parameters**
- $\bar{x}$ is an **estimate** of μ
- $\bar{x}$ is a realisation of the **estimator** $\bar{X}$
- **Standard Error (SE)** of the estimate (a number) is defined as the SD of the estimator

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

  which is how much $\bar{X}$ fluctuates around μ (a number) estimated from the data
- Estimate of σ

- Biased estimate of σ^2

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

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

- Unbiased estimate of σ^2 (preferred)

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

$$E(s^2) = \sigma^2$$

How to estimate μ?

- μ is estimated by $\bar{x}$
- Error in $\bar{x}$ is measured by the SE:

$$\text{SD}(\bar{x}) = \frac{\sigma}{\sqrt{n}}$$

which is **estimated** by $\frac{s}{\sqrt{n}}$ since σ is unknown
- **Conclusion**: μ is estimated as $\bar{X}$, give or take $\frac{s}{\sqrt{n}}$

$$\text{SE estimated by } \frac{s}{\sqrt{n}} = \frac{\sqrt{\frac{n}{n-1}} \times \text{SD}}{\sqrt{n}}$$

$$\text{where SD } = \hat{\sigma}$$

How to estimate p?

- $\hat{p}$ is the estimator of p

$$E(\hat{p}) = p$$

$$\text{var}(\hat{p}) = \frac{\sigma^2}{n} = \frac{p(1-p)}{n}$$

$$\text{SE} = \text{SD}(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}$$

which is **estimated** by realisations of $\hat{p}$

2.3 Interval estimation

2.3.1 Definitions

- For sufficiently large n,

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

- The p-quantile of $Z \sim \mathcal{N}(0,1)$ is the number q such that

$$\Phi(q) = \Pr(Z \leq q) = p$$
$$q = \Phi^{-1}(p)$$

```r
q <- qnorm(p)
p <- pnorm(q)
```

- For $0 < p < 0.5$, let z_p be such that

$$\Pr(Z > z_p) = p$$

$$z_p = \Phi^{-1}(1-p)$$

In other words, $z_p = (1-p)$-quantile of Z

2.3.2 CI Estimation

- For large n,

$$\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

$$\Pr\left(-z_{\frac{\alpha}{2}} \leq \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq z_{\frac{\alpha}{2}}\right) \approx 1-\alpha$$

$$\Pr\left(\bar{X} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right) \approx 1-\alpha$$

where the above, $\left(\bar{X} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right)$ is a random interval. Realisation $\bar{x}$ of $\bar{X}$ gives the realised interval
- $(1-\alpha)$-CI for μ is of the form

$$\left(\text{estimate} - z_{\frac{\alpha}{2}}\text{SE}, \text{estimate} + z_{\frac{\alpha}{2}}\text{SE}\right)$$

2.3.3 Exact CI

- Let $t_{\frac{\alpha}{2},n-1}$ be the number such that

$$\Pr(t_{n-1} > t_{\frac{\alpha}{2},n-1}) = \alpha/2$$

- **[Important]** Exact CI only works if $X \sim \mathcal{N}(\mu, \sigma^2)$ and x_i's are realisations from IID Normal Distribution
  * CI is exact means that $\Pr(\mu$ is within the interval$)$ is exactly $1-\alpha$
- $(1-\alpha)$-CI for μ is

$$\left(\bar{x} - t_{\frac{\alpha}{2},n-1}\frac{s}{\sqrt{n}}, \bar{x} + t_{\frac{\alpha}{2},n-1}\frac{s}{\sqrt{n}}\right)$$

2.4 Bias in Survey

Famous example: US presidential election survey conducted by *Literary Digest* in 1936

2.4.1 Bias in Measurement

- $x_1, \dots, x_n$ are realisations of random draws $X_i, \dots, X_n$ from a population with mean $\mu + b$ and variance σ^2
- $\text{SE} = \sigma/\sqrt{n}$ measures how far $\bar{x}$ is from $E(\bar{X}) = \mu + b$
- **Definition of Bias**

$$\text{Bias of estimate} = \text{E(estimator)} - \text{parameter}$$

- MSE

$$\text{E}(\bar{X}-\mu)^2 = \text{var}(\bar{X}) + \{\text{E}(\bar{X}) - \mu\}^2$$
$$\text{MSE} = \text{SE}^2 + \text{bias}^2$$

However μ is unknowable, hence it is not possible to remove bias unless we make very careful observations

3 Parameter Estimation

Assuming data $x_1, \dots, x_n$ are realisations of IID RV's $X_1, \dots, X_n$ with density $f(x|\theta)$, estimate θ.
The parameter θ lies in $\Theta \subseteq \mathbb{R}$ where Θ is the parameter space
How to estimate θ from realisations $x_1, \dots, x_n$?

1. Method of moments
2. Method of maximum likelihood

3.1 Method of moments

Let $\hat{\theta}$ be an estimator for θ.
The k-th moments of an RV X is

$$\mu_k = E(X^k)$$

$$\frac{1}{n}\sum_{i=1}^{n} x_i^k$$

is a realisation of $\hat{\mu_k}$ and is used as estimate for μ_k

$$\hat{\theta} = g(\hat{\mu_1}, \dots, \hat{\mu_q})$$

is an estimate for θ *e.g.* for Normal RV,

$$g : \begin{bmatrix} x \\ y \end{bmatrix} \to \begin{bmatrix} x \\ y - x^2 \end{bmatrix}$$

3.2 Monte Carlo Approximation

Needed if formula for θ is complicated/hard to compute the value of its expectation

Rough Steps:

1. Estimate parameters θ using MOM/MLE
2. Generate n realisations $x_1, x_2, \dots, x_n$ using the estimated parameters and distribution
3. From these n realisations, estimate parameters again, these are realisations of $\hat{\theta}^*$
4. Repeat steps 2 and 3 m times until we get m realisations of parameters θ

$$\mathsf{SE} = \mathsf{SD}(\hat{\theta}) \approx \mathsf{SD}(\hat{\theta}^*)$$

$$\mathsf{Bias} = \mathsf{E}(\hat{\theta}) - \theta \approx \mathsf{E}(\hat{\theta}^*) - \theta_{\text{est.}}$$

5. Finally, θ is around $\theta_{\text{est.}} -$ Bias $\pm$ SE, and the fitted distribution + parameter is called a **statistical model** for the event in question

Note that as $n \to \infty$, $\mathsf{E}(\hat{\theta}^*) \to \theta_{\text{est}} \Rightarrow$ Bias$\to 0$, $\mathsf{E}(\hat{\theta}) \to \theta$.

- Thus, it is **asymptotically unbiased**
- Every MOM estimator is consistent, it goes to the parameter as $n \to \infty$

3.3 Maximum Likelihood Method

Let $x_1, \dots, x_n$ be realisations of IID RV's $X_1, \dots, X_n$ with density/mass function $f(x|\theta)$

$$L(\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

$$l(\theta) = \log L(\theta) = \sum_{i=1}^{n} \log f(x_i|\theta)$$

Find the value of θ that maximises the likelihood

3.3.1 Multinomial Data

$$L(p_1, \dots, p_r) = p_1^{x_1} \dots p_r^{x_r} \times c$$
$$l(p_1, \dots, p_r) = x_1 \log p_1 + \dots + x_r \log p_r + \log c$$

Since $p_1 + \dots + p_r = 1$, differentiating l does not work since it's constrained, hence we use the **Lagrangian** function and treating $p_1, \dots, p_r, \lambda$ as if they are unconstrained

$$\mathcal{L}(p_1, \dots, p_r, \lambda) = x_1 \log p_1 + \dots + x_r \log p_r + \lambda(p_1 + \dots + p_r - 1)$$

3.3.2 Genetics

Chromosomes come in pairs, one from each parent
Locus a subsequence on a chromosome
Alleles different versions of bases at a locus
Genotype an unordered pair of alleles

- Given k different alleles, we can construct $k(k+1)/2$ different genotypes
- Given the genotype proportions, we can calculate the allele proportions
- Given the allele proportions, we can calculate the genotype proportions

Mendel's Laws of inheritance

- The maternal allele is randomly chosen from her two alleles; similarly for the paternal allele
- The two choices are independent

Hardy-Weinberg Equilibrium: A population is in HWE at a locus if the genotype proportions are

$$f(a_i a_j) = \begin{cases} p_i^2 & i = j \\ 2p_i p_j & i \neq j \end{cases}$$

where p_i is the proportion of allele a_i (assumption: random mating, no mutation, no migration)

3.4 Large-Sample Variance of ML Estimator

Let X have density $f(x|\theta)$, $\theta \in \Theta \subset \mathbb{R}^p$. The Fisher information is the $p \times p$ matrix

$$\mathcal{I}(\theta) = -\mathsf{E}\left[\frac{d^2 \log f(X)}{d\theta^2}\right]$$

with (i, j) entry

$$-\mathsf{E}\left[\frac{\partial^2 \log f(X)}{\partial\theta_i \partial\theta_j}\right]$$

$$= -\int_{-\infty}^{\infty} \frac{\partial^2 \log f(x)}{\partial\theta_i \partial\theta_j} f(x) dx \text{ or } -\sum_x \frac{\partial^2 \log f(x)}{\partial\theta_i \partial\theta_j} f(x) dx$$

As $n \to \infty$,

$$\mathsf{var}(\hat{\theta}_n) \approx \frac{\mathcal{I}(\theta)^{-1}}{n}$$

3.4.1 Joint Density

IID $X_1, \dots, X_n$ with density $f(x|\theta)$ can be regarded as a sample from $\mathbf{X} = (X_1, \dots, X_n) \in \mathbb{R}^n$ with joint density

$$g(\mathbf{X}|\theta) = f(X_1|\theta) \cdots f(X_n|\theta)$$

The information in $\mathbf{X}$ is

$$- - \mathsf{E}\left[\frac{d^2 \log g(\mathbf{X})}{d\theta^2}\right] = n\mathcal{I}(\theta)$$

where $\mathcal{I}(\theta)$ is the information in any one of the X's

3.4.2 Multinomial

Let **X** $\sim$ Multinomial$(n, \mathbf{p}(\theta))$ where

$$\mathbf{p}(\theta) = (p_1(\theta), \dots, p_r(\theta))$$

$$\theta \in \Theta \subset \mathbb{R}^k, 1 \leq, k \leq r-1$$

Then,

$$\log f(\mathbf{X}) = \sum_{i=1}^{r} X_i \log p_i$$

(i, j) entry of $\mathbb{I}(\theta)$:

$$\frac{n}{p_i} + \frac{n}{p_r}, i = j$$
$$\frac{n}{p_r}, i \neq j$$

3.5 Distribution of MLE

As $n \to \infty$, the distribution of

$$\sqrt{n\mathcal{I}(\theta)}(\hat{\theta}_n - \theta)$$

converges to $\mathcal{N}(\mathbf{0}, \mathbf{I}_p)$.
For large n,

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

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

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

$$1 - \alpha \approx \Pr(\text{estimate} - z_{\frac{\alpha}{2}}\text{SE} \leq \theta \leq \text{estimate} + z_{\frac{\alpha}{2}}\text{SE})$$

Where estimate is drawn using MLE from data, and SE is drawn using the estimate and Fischer information

3.5.1 Asymptotic Normality

Let $\hat{\theta}$ be the ML estimator of θ.

- For any strictly decreasing/increasing function $h : \Theta \to \mathbb{R}$, $h(\hat{\theta})$ is also the ML estimator of $h(\theta)$.
- For large n, $h(\hat{\theta})$ is approximately normal (**asymptotically normal**)

3.6 ML vs MOM

- Both ML and MOM are **consistent**: bias goes to 0 as $n \to \infty$
- ML is better (smaller bias and SE) because it uses all info contained in the density function, whereas MOM uses only sample moments to estimate parameters
- ML estimators have **asymptotic properties**: as $n \to \infty$, SE can be estimated without Monte Carlo and so a good CI for the parameter is available
- MOM estimators may not be asymptotically normal so it is more difficult to construct a CI, but it is easier to compute so is sometimes useful

4 Goodness-of-fit

4.1 Pearson's X^2 Test

Let $(X_1, \dots, X_r) \sim$ Multinomial(n,**p**) with n, r fixed. Then the set of all possible distributions of **p** is:

$$\Omega = \left\{(p_1, \dots, p_r) : p_i > 0, \sum_{i=1}^{r} p_i = 1\right\}$$

Consider a subset Ω_0 where **p** depends on $\theta \in \Theta \subset \mathbb{R}^k, k, r-1$

$$\Omega_0 = \{(p_1(\theta), \dots, p_r(\theta)) : \theta \in \Theta\}$$

Now we want to judge if $\mathbf{p} \in \Omega_0$ given realisations $(x_1, \dots, x_r)$ (in other words, is **p** a function which takes in a k-dimensional vector θ)

- Assuming $(X_1, \dots, X_r) \sim$ Multinomial$(n, \mathbf{p}(\theta)), \theta \in \Theta \subset \mathbb{R}^k, k < r-1$
- $\hat{\theta}$ is the ML estimator of θ
- $n\mathbf{p}(\hat{\theta})$ is the random expected counts
- **Chi-square statistic**:

$$X^2 = \sum_{i=1}^{r} \frac{(X_i - np_i(\hat{\theta}))^2}{np_i(\hat{\theta})} = \sum \frac{(O-E)^2}{E}$$

  [Theorem] As $n \to \infty$, the distribution of X^2 converges to χ^2_{r-1-k}
  Note that <u>k can be 0</u>, in the case of assuming fair die where there is no parameter to estimate (when the properties are equal)

<u>Steps for X^2 goodness-of-fit test</u>

1. Let $H_0 : \mathbf{p} \in \Omega_0$
2. Let $H_1 : \mathbf{p} \in \Omega_1$
3. Substituting each X_i by x_i (the observed realisations) and $\hat{\theta}$ by the ML estimate (to get the expected counts), we get a realisation x^2 of X^2
4. The P-value: (Calculated assuming H_0)

$$\Pr(X^2 \geq x^2) \approx \Pr(\chi^2_{r-1-k} \geq x^2)$$

   The smaller it is, the more suspicious we are of H_0 (more likely to reject H_0)
   The bigger it is, we are more likely to accept H_0

4.2 Likelihood Ratio

Assuming multinomial, Maximum of likelihood $L(\mathbf{p}) = \prod_{i=1}^{r} p_i^{X_i}$ over Ω, happens when there is no restriction, do MLE as usual

$$L_1 = L(\hat{p}) = \prod_{i=1}^{r}\left(\frac{X_i}{n}\right)^{X_i}$$

Maximum of likelihood $L(\theta) = \prod_{i=1}^{r} p_i(\theta)^{X_i}$ over Ω_0

$$L_0 = L(\hat{\theta}) = \prod_{i=1}^{r} p_i(\hat{\theta})^{X_i}$$

Note that $L_0/L_1 \geq 1$, the larger the ratio, the more we doubt H_0

$$2\log\left(\frac{L_1}{L_0}\right) = G$$

$$G = 2\sum_{i=1}^{r} X_i \log\left(\frac{X_i}{np_i(\hat{\theta})}\right)$$

4.2.1 LR goodness-of-fit test

Assumptions
- n IID RV's density defined by $\theta \in \Omega$ with k independent parameters
- L_1: maximum likelihood value over Ω
- L_0: maximum likelihood value over Ω_0 with $k_0 < k_1$ independent parameters

Theorem: Suppose $\theta \in \Omega_0$ (Assume H_0 is true). As $n \to \infty$, the distribution of

$$G = 2\log\left(\frac{L_1}{L_0}\right)$$

converges to $\chi^2_{k_1-k_0}$ **LR goodness-of-fit test**
1. $H_0 : \theta \in \Omega_0$
2. $H_1 : \theta \in \Omega_1$
3. L_0 and L_1 are the maximum likelihood values under Ω_0 and Ω_1
$$g = 2\log\left(\frac{L_0}{L_1}\right)$$
is a realisation of G
4. The P-value is calculated with distribution of G under H_0
$$\Pr(G \geq g) \approx \Pr(\chi^2_{k_1-k_0} \geq g)$$

4.2.2 Conclusion

- The LR test assumes the larger model is valid, and does not assess its goodness-of-fit
- P-value is not a probability that H_0 is true, P-value is computed assuming H_0 is true

4.3 Poisson Dispersion Test

For Poisson, if var is more or less the same as mean, then it can fit well. But if var is $\gg$ mean then need to find new distribution or the data might come from two or more different RV's
1. $H_1 : \theta \in \Omega$: For $i = 1, \ldots, n$, $X_i \sim$ Poisson(λ_i) and each are independent
$$l(\lambda_1, \ldots, \lambda_n) = \sum_{i=1}^{n} X_i \log \lambda_i - \sum_{i=1}^{n} \lambda_i$$
When $l(\lambda_1, \ldots, \lambda_n)$ is maximum, $\hat{\lambda} = X_i$, so maximum likelihood under $\Omega : l_1 = \sum_{i=1}^{n} X_i \log X_i - \sum_{i=1}^{n} X_i$
2. $H_0 : \theta \in \Omega_0$: Every $\lambda_i = \lambda$
$$l(\lambda) = \sum_{i=1}^{n} X_i \log \lambda - n\lambda$$
Maximum likelihood is achieved when $\hat{\lambda} = \bar{X}$ under $\Omega_0 : l_0 = \sum_{i=1}^{n} X_i \log \bar{X} - n\bar{X}$
3. Calculate P-value
$$G = 2\sum_{i=1}^{n} X_i \log\left(\frac{X_i}{\bar{X}}\right) \approx \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{\bar{X}}$$
Suppose every $\lambda_i = \lambda$. For large n, $G \sim \chi^2_{n-1}$ approximately

5 Useful Results

5.1 Algebra

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

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

5.2 Procedures

Framework for statistical inference:
1. Parameter is a simple function of the population, real or hypothetical
2. Data are realisations of IID RV's (if $n \ll N$)
3. Estimate is a realisation of an estimator, whose SD is the SE. For large n, can construct CI.
4. MSE = SE^2 + bias^2

5.3 Multivariable Calculus

- Use Hessian matrix to calculate partial derivatives/-maximum points, and $|H| > 0$