1 Probability

- Baye's theorem $P(B) = P(B \mid A)P(A) + P(B \mid A^C)P(A^C)$.
- Baye's first formula $P(B) = \sum P(B \mid A_i)P(A_i)$.
- Baye's second formula $P(A_i \mid B) = \frac{P(B|A_i)P(A_i)}{\sum P(B|A_j)P(A_j)}$.

2 Distributions

2.1 Binomial distribution

- Total number of successes in n Bernoulli trials.
- PDF: $P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$.
- MGF: $M_x(t) = (1 - p - pe^t)^n$.
- $\mathrm{E}(X) = np$, $\mathrm{Var}(X) = np(1-p)$, $I(\theta) = \frac{n}{p(1-p)}$.
- When n large, p small, np moderate, $\mathrm{Bin}(n,p) \approx \mathrm{Po}(np)$.

2.2 Negative binomial distribution

- Number of independent Bernoulli trials performed until r successes.
- PDF: $P(X = k) = \binom{k-1}{r-1} p^r (1-p)^{k-r}$.
- MGF: $M(t) = \left(\frac{1-p}{1-pe^t}\right)^r$.
- $\mathrm{E}(X) = r(1-p)/p^2$, $I(\theta) = \frac{r}{p(1-p)^2}$.

2.3 Geometric distribution

- Infinite Bernoulli trials, total number of trials up to and including the first success.
- PDF: $P(X = k) = p(1-p)^{k-1}$.
- CDF: $1 - (1-p)^k$.
- MGF: $M_x(t) = \frac{pe^t}{1-(1-p)e^t}$.
- $\mathrm{E}(X) = 1/p$, $\mathrm{Var}(X) = (1-p)/p^2$.

2.4 Poisson distribution

- PDF: $P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}$
- MGF: $M_x(t) = e^{\lambda(e^t - 1)}$.
- $\mathrm{E}(X) = \lambda$, $\mathrm{Var}(X) = \lambda$, $I(\theta) = 1/\lambda$.

2.5 Exponential distribution

- PDF: $f(x) = \lambda e^{-\lambda x}$ for $x \geq 0$.
- CDF: $F(x) = \int_{-\infty}^{x} f(u)\,\mathrm{d}u = 1 - e^{-\lambda x}$ for $x \geq 0$.
- $\mathrm{E}(X) = 1/\lambda$, $\mathrm{Var}(X) = 1/\lambda^2$, $I(\theta) = 1/\lambda^2$.

2.6 Gamma distribution

- PDF: $f(x) = \frac{\lambda e^{-\lambda x}(\lambda x)^{\alpha-1}}{\Gamma(\alpha)}$, where $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x}\,\mathrm{d}x$.
- MGF: $(1 - t/\lambda)^{-\alpha}$.
- $\mathrm{E}(X) = \alpha/\lambda$, $\mathrm{Var}(X) = \alpha/\lambda^2$.
- $\Gamma(1, \lambda) = \mathrm{Exp}(\lambda)$.

2.7 Normal distribution

- PDF: $f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$.
- MGF: $M_x(t) = \exp\left(\mu t + \frac{\sigma^2 t^2}{2}\right)$.
- $\mathrm{E}(X) = \mu$, $\mathrm{Var}(X) = \sigma^2$.

  Let $X_1, \ldots, X_n$ be sampled from a normal distribution. Define the sample mean and variance

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

Then $\mathrm{E}(\bar{X}) = \mu$, $\mathrm{Var}(\bar{X}) = \sigma^2/n$. Furthermore X and S^2 are independent.

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1} \tag{2}$$

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1} \tag{3}$$

2.8 Chi square distribution

If Z is standard normal, then $U = Z^2$ is chi-square with 1 dof. If U_i are chi-square with 1 dof, then $V = U_1 + \cdots + U_n$ is chi-square with n dof.

- $\mathrm{E}(V) = n$, $\mathrm{Var}(V) = 2n$.
- MGF: $M(t) = (1 - 2t)^{-n/2}$.
- $\chi^2_n = \Gamma(\alpha = \frac{n}{2}, \lambda = \frac{1}{2})$
- $\chi^2_m + \chi^2_n = \chi^2_{m+n}$.

2.9 t distribution

If Z is standard normal and $U \sim \chi^2_n$, then $Z/\sqrt{U/n}$ is a t distribution with n dof.

- PDF: $f(t) = \frac{\Gamma[(n+1)/2]}{\sqrt{n\pi}\Gamma(n/2)}(1 + t^2/n)^{-(n+1)/2}$.

2.10 F distribution

$U \sim \chi^2_m$ and $V \sim \chi^2_n$ then $W = \frac{U/m}{V/n}$ is a F distribution with m and n dof.

3 Random variables

Let X have PDF f_X CDF F_X, and $Y = aX + b$, then

- $F_Y(y) = P(Y \leq y) = F_X(\frac{y-b}{a})$,
- $f_y(y) = \frac{\mathrm{d}}{\mathrm{d}y} F_Y(y) = \frac{1}{a} f_X(\frac{y-b}{a})$.
- Let U be uniform on $[0,1]$ and $X = F^{-1}(U)$. Then the CDF of X is F.

4 Extrema and order statistics

$X_1, \ldots, X_n$ have CDF F and density f. Let U be their max and V be their min.

- $F_U(u) = P(U \leq u) = [F(u)]^n$, $f_U(u) = nf(u)[F(u)]^{n-1}$.
- $F_V(v) = 1 - [1 - F(v)]^n$, $f_V(v) = nf(v)[1 - F(v)]^{n-1}$.
- $f_k(x) = \frac{n!}{(k-1)!(n-k)!} f(x)[F(x)]^{k-1}[1 - F(x)]^{n-k}$.

5 Expectation value and variance

Let $Y = g(X)$. Then $\mathrm{E}(Y) = \sum g(x)p(x)$ or $\mathrm{E}(Y) = \int_{-\infty}^{\infty} g(x)f(x)\,\mathrm{d}x$ (can generalise for joint variables).

- If X is nonnegative continuous, $\mathrm{E}(X) = \int_0^\infty 1 - F(x)\,\mathrm{d}x$.
- If X and Y are independent then $\mathrm{E}[g(X)h(Y)] = E[g(X)]E[h(Y)]$.
- $\mathrm{E}[a + bX] = a + b\,\mathrm{E}[X]$.

Define $\mathrm{Var}(X) = \sum(x_i - \mu)2p(x_i)$ or $\mathrm{Var}(X) = \int_{-\infty}^{\infty}(x-\mu)^2 f(x)\,\mathrm{d}x$. Define $\mathrm{Cov}(X,Y) = E[(X - \mu_X)(Y - \mu_Y)]$.

- $\mathrm{Var}(a + bX) = b^2\,\mathrm{Var}(X)$.
- $\mathrm{Var}(X) = E(X^2) - E(X)^2$.

6 Moment generating functions

The MGF of X is defined as $M(t) = E[e^{tX}]$.

- If the MGF exists on an open interval containing 0, then it uniquely determines the probability distribution.
- If X and Y are independent, then $M_{X+Y}(t) = M_X(t)M_Y(t)$.

7 Delta method

Perform Taylor expansion about μ_X:

$$Y = g(X) \approx g(\mu_X) + (X - \mu_X)g'(\mu_X) + \frac{1}{2}(X - \mu_X)^2 g''(\mu_X). \tag{4}$$

$$\mathrm{E}(Y) \approx g(\mu_X) + \frac{1}{2}\sigma_X^2 g''(\mu_X) \tag{5}$$

$$\mathrm{Var}(Y) \approx \sigma_X^2 [g'(\mu_X)]^2 \tag{6}$$

8 Central Limit Theorem

Theorem 8.1. *Let $X_1, \ldots$ be IID with mean 0 and variance σ^2. Let $S_n = \sum_i^n X_i$. Then*

$$\lim_{n\to\infty} P(\frac{S_n}{\sigma\sqrt{n}} \leq x) = \Phi(x). \tag{7}$$

where Φ is the CDF of the normal distribution.

9 Parameter estimation

9.1 Method of moments

Define the k-th *sample* moment as $\hat{\mu}_k = \frac{1}{n}\sum_i X_i^k$. Suppose we want to estimate θ_1 and θ_2. Express θ_1 and θ_2 in terms of the actual moments:

$$\theta_1 = f_1(\mu_1, \mu_2) \qquad \theta_2 = f_2(\mu_1, \mu_2) \tag{8}$$

then the method of moments estimates are

$$\hat{\theta}_1 = f_1(\hat{\mu}_1, \hat{\mu}_2) \qquad \hat{\theta}_2 = f_2(\hat{\mu}_1, \hat{\mu}_2) \tag{9}$$

We can use bootstrap to simulate N samples of size n from the distribution with $\hat{\theta}_1$ and $\hat{\theta}_2$. For each sample, calculate MOM estimates θ_1^* and θ_2^*. Use N values of $*$ to approximate sampling dist.

9.2 Maximum likelihood estimate

If X_i are iid, then define likelihood $lik(\theta) = \prod f(X_i \mid \theta)$. Then find maxima for $l(\theta) = \log(lik(\theta))$. Bootstrap can also be used. Just change MOM to MLE above.

Suppose now $X_1, \ldots, X_m$, the counts in cells $1, \ldots, m$, follow a multinomial distribution with cell probabilities $p_1, \ldots, p_m$ that we want to estimate. Use Lagrange multiplier

$$L(p_1, \ldots, p_m, \lambda) = \log n! - \sum_{i=1}^m \log x_i! + \sum_{i=1}^m x_i \log p_i + \lambda\left(\sum_{i=1}^m p_i - 1\right) \tag{10}$$

solve $\boldsymbol{\nabla} L = 0$.

Let θ_0 be the true value.

- Under appropriate conditions, the MLE is consistent, i.e. $\hat{\theta}$ converges to θ_0 in probability.
- Under appropriate conditions,
$$I(\theta) = E\left[\frac{\partial}{\partial\theta}\log f(X \mid \theta)\right]^2 = -E\left[\frac{\partial^2}{\partial\theta^2}\log f(X \mid \theta)\right] \tag{11}$$
- Under approriate conditions, $\sqrt{nI(\theta_0)}(\hat{\theta} - \theta_0)$ tends to standard normal.

Confidence intervals:

- For μ, $\bar{X} \pm \frac{S}{\sqrt{n}} t_{n-1}(\alpha/2)$.
- For σ^2. $\left(\frac{n\hat{\sigma}^2}{\chi^2_{n-1}(\alpha/2)}, \frac{n\hat{\sigma}^2}{\chi^2_{n-1}(1-\alpha/2)}\right)$.
- Approximate CI for θ_0: $\hat{\theta} \pm \frac{z(\alpha/2)}{\sqrt{nI(\hat{\theta})}}$.
- Use bootstrap. Generate B samples from a dist. with $\hat{\theta}$ and for each sample make estimate θ^*. Approximate distribution $\hat{\theta} - \theta_0$ by $\theta^* - \hat{\theta}$. Use quantiles to make an approximate CI, $P(\hat{\theta} - \bar{\delta} \leq \theta_0 \leq \hat{\theta} - \underline{\delta}) = 1 - \alpha$.

9.3 Bayesian approach

If we have prior distribution $f_\Theta(\theta)$, the distribution of Θ given the data X is the posterior distribution:

$$f_{\Theta|X}(\theta \mid x) = \frac{f_{x|\theta}(x \mid \theta) f_\Theta(\theta)}{\int f_{x|\theta}(x \mid \theta) f_\Theta(\theta)\,\mathrm{d}\theta}. \tag{12}$$

9.4 Efficiency

Mean square error is also $\mathrm{Var}(\hat{\theta}) + (E(\hat{\theta}) - \theta_0)^2$. If an estimate is unbiased, $E(\hat{\theta}) = \theta_0$ and MSE becomes $\mathrm{Var}(\hat{\theta})$. Efficiency is defined as the ratio of variances.

Theorem 9.1 (Carmer-Rao inequality)**.** *Under appropriate conditions, if T is an unbiased estimate of θ, then $\mathrm{Var}(T) \geq 1/(nI(\theta))$.*

9.5 Sufficiency

A statistic $T(X_1, \ldots, X_n)$ is sufficient for θ is the conditional distribution of $X_1, \ldots, X_n$ given $T = t$ does not depend on θ.

A statistic $T(X_1, \ldots, X_n)$ is sufficient for θ iff the joint probability function factors $f(x_1, \ldots, x_n \mid \theta) = g(T(x_1, \ldots, x_n), \theta)h(x_1, \ldots, x_n)$.

Theorem 9.2 (Rao-Blackwell)**.** *If T is sufficient for θ, let $\tilde{\theta} = E(\hat{\theta} \mid T)$. Then $E(\tilde{\theta} - \theta)^2 \leq E(\hat{\theta} - \theta)^2$.*

10 Hypothesis testing

- Rejecting H_0 when it is true is called type I error, its probability is the significance level, denoted α.
- Accepting H_0 when it is false is called type II error, its probability is β.
- The probability that H_0 is rejected when it is false is called the power of the test, given by $1 - \beta$.
- Likelihood ratio or test statistic: $P(x \mid H_0)/P(x \mid H_1)$.
- Simple hypotheses completely specify the probability distribution.

Theorem 10.1 (Neyman-Pearson lemma)**.** *Suppose H_0 and H_1 are simple hypotheses and the test that rejects H_0 whenever the likelihood ratio is less that c has significance level α. Then any other test which has significance level leq α has power leq that of the likelihood ratio test.*

Theorem 10.2. *Suppose that for every $\theta_0 \in \Theta$ there is a test at level α of the hypothesis that $\theta = \theta_0$. Denote the acceptance region as $A(\theta_0)$. Then the set $C(X) = \{\theta \mid X \in A(\theta)\}$ is a $1 - \alpha$ confidence region for θ.*

Theorem 10.3. *Suppose that $C(X)$ is a $1 - \alpha$ confidence region for θ, that is, for every θ_0, $P[\theta_0 \in C(X) \mid \theta = \theta_0] = 1 - \alpha$. Then an acceptance region for a test at level α of the hypothesis $\theta = \theta_0$ is $A(\theta_0) = \{X \mid \theta_0 \in C(X)\}$.*

10.1 Generalised likelihood ratio tests

Suppose hypotheses H_0 has parameter space ω_0 and H_1 has parameter space ω_1, and let $\Omega = \omega_0 \cup \omega_1$. We like to use the test statistic $\Lambda = \frac{\max_{\theta\in\omega_0}(lik(\theta))}{\max_{\theta\in\Omega}(lik(\theta))}$. The rejection threshold is chosen such that $P(\Lambda \leq \lambda_0 \mid H_0) = \alpha$.

For the multinomial distribution, the likelihood ratio is given by $\Lambda = \prod_i^m \left(\frac{p_i(\hat{\theta})}{\hat{p}_i}\right)^{x_i}$. Pearson's statistic is more commonly used to test goodness of fit: $\chi^2 = \sum_i^m \frac{[x_i - np_i(\hat{\theta})]^2}{np_i(\hat{\theta})}$.