Probability

Concepts

Independence

Informal:

$$P(E) = P(E|F)$$

Formal:

$$P(E,F) = P(E) \cdot P(F)$$

Law of Total Probability

$$P(E) = P(E|F) \cdot P(F) + P(E|F^c) \cdot P(F^c)$$

Bayes' Rule

$$\begin{aligned} P(E|F) &= \frac{P(E,F)}{P(F)} \\ &= \frac{P(F|E) \cdot P(E)}{\int_e P(F|e) \cdot P(e)} \end{aligned}$$

Expectation & Variance & Covariance

1. Expectation: Mean, average value

$$E[X] = \int_x x \cdot P(x) dx$$

$$E[g(x)] = \int_x g(x) \cdot P(x) dx$$

```
Linearity of Expectation:
```

$$E[X + Y] = E[X] + E[Y]$$

$$E[aX + b] = aE[X] + b$$

3. Variance: "Distance" from mean

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

```
Quadraticity of Variance:
```

$$Var[X+Y] = Var[X] + Var[Y] + Cov(X,Y)$$

$$Var[aX+b] = a^2 Var[X]$$

3. Covariance: Relation

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

```
Affine Transformation
```

$$Cov(aX+b, cY+d) = acCov(X,Y)$$

4. Correlation Coefficient: Ensure the correlation is not affected by the scale of the variables.

$$\rho_{XY} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

Distributions

Approximation

- Stirling's Approximation: $n! \approx \sqrt{2\pi n}(\frac{n}{e})^n$
- $(1-\frac{1}{n})^n = \frac{1}{e};\ (1+\frac{1}{n})^n = e$

Discrete Distributions

Bernoulli Distribution

Meaning: A single trail succeeds with probability p, and fails with probability $1-p$. Denoted as B_p.

Formula:

- PMF: $P_{X|\theta}(x|p) = p^x(1-p)^{1-x}$
- CDF: No special formula

Statistics: $E[X] = p, Var[X] = pq$

Binomial Distribution

Meaning: Number of success in n **independent** Bernoulli trials with success probability p. Denoted as $B_{p,n}$.

Formula:

- PMF: $P_{X|theta}(x|p,n) = \binom{n}{x}p^x(1-p)^{n-x}$
- CDF: No special formula

Statistics: $E[X] = np, Var[X] = npq$

Properties: Can be approximated using **Poisson Distribution** and **Gaussian Distribution**.

Poisson Distribution

Meaning: Approximation of **Binomial Distribution** when $n \gg 1 \gg p$. Parameter $\lambda = np$, represents the occurence rate in unit time.

Formula:

- PMF: $P_{X|theta}(x|\lambda) = e^{-\lambda}\frac{\lambda^k}{k!}$
- CDF: No special formula

Statistics: $E[X] = \lambda, Var[X] = \lambda$

Properties: For independent $X,\ Y$; $X \sim P_\lambda$; $Y \sim P_\gamma$; $(X+Y) \sim P_{\lambda+\gamma}$

Geometric Distribution

Meaning: Obtaining the first success in a series of **Bernoulli trials**.

Formula:

- PMF: $P_{X|\theta}(n|p) = p \cdot q^{n-1}$
- CDF: $P(n > N) = q^N$; $P(n \leq N) = 1 - q^N$

Statistics: $E[X] = \frac{1}{p}, Var[X] = \frac{q}{p^2}$

Properties:

- Estimate upper bound w/o p: $P(n) \leq \frac{1}{e(n-1)}$

- Memoryless: $P(n+m|X>n)=P(m)$; Geometric->Memoryless, Memoryless->Geometric.
- $E[X|X>n]=E[n+X]$

Application:

Coupon Collector Problem: Assume we have 3 different types of coupon, with uniform probability. How many coupons do we need to sample to collect all three coupons?

1. Denote the number needed to obtain the first, second, third coupon as X_1, X_2, X_3, these three random variables follow **geometric distribution** $G_1, G_{\frac{2}{3}}, G_{\frac{1}{3}}$. Then, the total number of sample needed can be calculated as $E[X_1+X_2+X_3]=E[X_1]+E[X_2]+E[X_3]=1+\frac{3}{2}+\frac{3}{1}=5.5$

Continuous Distributions

Continuous function transformation: $X\sim f_X(x), Y=g(X), f_Y(y)=\sum_{x=g^{-1}(y)}\frac{f_X(x)}{|g'(x)|}$, substitute all the x with $g^{-1}(y)$. Unit length Δx around x, but $|g'(x)\Delta x|$ around y.

Exponential Distribution

Meaning: Obtaining the first success at time x. Geometric + Poisson = Exponential. $\lambda=np$, using the PMF of **Geometric**, $P(X>x)=(1-\frac{\lambda}{n})^{nx}=(1-\frac{\lambda}{n})^{\frac{n\lambda x}{\lambda}}=e^{-\lambda x}$.

Formula:

- PDF: $P_{X|\theta}(x|\lambda)=\lambda e^{-\lambda x}$
- CDF: $F(X)=1-e^{-\lambda x}$

Statistics: $E[X]=\frac{1}{\lambda}, Var[X]=\frac{1}{\lambda^2}$

Properties:

- $X\sim E_\lambda,\ Y=aX,\ Y\sim E_{\frac{\lambda}{a}}$
- Memoryless: $P(X<a+b|X\geq a)=P(X<b),\ E[X|X\geq a]=E[X]+a.$

Gaussian Distribution

Formula:

- PDF: $P_{X|\theta}(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
- CDF: **Z-Table**

Statistics: $E[X] = \mu, Var[X] = \sigma^2$

Properties:

- $X \sim N(\mu, \sigma^2),\ Y = aX + b,\ Y \sim N(a\mu + b, a^2\sigma^2)$. Any Gaussian distribution can be transformed to Standard Gaussian Distribution Z.
- Sum of independent Gaussian is still Gaussian: $G_1 \sim (\mu_1, \sigma_1^2),\ G_2 \sim (\mu_2, \sigma_2^2),\ G_1 + G_2 = G_3,\ G_3 \sim (\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$
- De Moivre - Laplace Theorem: $B_{p,n}(k) \approx N_{np,npq}(k)$.
- Approximate **Binomial** with **Z-Table**: When σ is large enough, the Gaussian PDF can be estimated as linear in a small range, therefore, $B_{p,\,n}(k) \approx N_{np,\,npq}(k) \approx \int_{k-0.5}^{k+0.5} N_{np,\,npq}(x)dx$. We can further estimate the sum of **Bernoulli** using this estimation: $\sum_{i\in[n,\,m]} B_{p,\,n}(i) \approx \int_{n-0.5}^{m+0.5} N_{np,\,npq}(x)dx$.

Inequalities

Markov Inequality

For non-negative r.v. X, $P[X \geq a\mu] \leq \frac{1}{a},\ P[X \geq a] \leq \frac{\mu}{a}$.

Chebyshev's Inequality

For any r.v. X, $P[|X - \mu| \geq a\sigma] \leq \frac{1}{a^2},\ P[|X - \mu| \geq a] \leq \frac{\sigma^2}{a^2}$.

- Proof:

Using the *Markov Inequality*

$$P[|X - \mu| \geq a] = P[(X - \mu)^2 \geq a^2] \leq \frac{\mu^{'}}{a^2}$$

where $\mu^{'} = E[(X - \mu)^2] = \sigma^2$.

- Application:

Weak Law of Large Number: $X_1, ..., X_n$ are independent & identically distributed samples, as n increases. the sample mean will converge to the actual mean of the distribution.

$$E[\bar{X}_i] = \frac{1}{n}\sum_i E[X_i] = E[X];\ Var[\bar{X}_i] = \frac{1}{n^2}\sum_i Var[X_i] = \frac{1}{n}Var[X]$$

Applying the *Chebyshev's Inequality*:

$$P[|\bar{X}_i - \mu| \geq \delta] \leq \frac{\sigma^2}{n\delta^2}$$

Chernoff Bound

Derived using *Markov Inequality*

$$P[X \geq a] = P[e^{tX} \geq e^{ta}] \leq \frac{E[e^{tX}]}{e^{ta}} = \frac{M_X(t)}{e^{ta}}$$

- Binomial: $P[X \geq (1+\delta)\mu] \leq e^{-\frac{\delta^2}{\delta+2}\mu};\ P[X \geq (1-\delta)\mu] \leq e^{-\frac{\delta^2}{2}\mu}$

Moment

- Raw moment: $E(X^n)$; Central moment $E[(X-\mu)^n]$.
- Moment generating function (MGF): $M_X(t) = E[e^{tX}]$.
- Properties:

$$M_{X+b}(t) = e^{bt}M_X(t)$$

$$M_{aX}(t) = M_X(at)$$

```
If$X$ and $Y$ are independent
```

$$M_{X+Y}(t) = M_X(t) \cdot M_Y(t)$$

Taylor Series: $e^y = 1 + y + \frac{y^2}{2!} + ... + \frac{y^n}{n!}$, then $M_X(t) = E[1 + tx + \frac{(tx)^2}{2!} + ... + \frac{(tx)^n}{n!}]$, therefore, $E[X^n] = M_X^{(n)}(t)$, the nth derivative of $M_X(t)$.

- Bernoulli: $M_X(t) = (1-p) + pe^t$
- Binomial: $M_X(t) = ((e^t - 1)p + 1)^n$

- Geometric: $M_X(t) = \frac{pe^t}{1-(1-p)e^t}$
- Poisson: $M_X(t) = e^{\lambda(e^t-1)}$
- Exponential: $M_X(t) = \frac{\lambda}{\lambda-t}$ if $\lambda < t$; else $M_X(t) = \infty$.
- Gaussian: $M_X(t) = e^{ut+\frac{\sigma^2 t^2}{2}}$; $E[X] = \mu$; $E[X^2] = \mu^2 + \sigma^2$.

Central Limit Theorem

Assume $\{X_1, ..., X_n\}$ i.i.d, when $n \to \infty$, $\bar{X} \sim G(\mu, \frac{\sigma^2}{n})$.

Statistic

Data sources: distribution / population; have certain attribute and parameter.

Sampling: i.i.d or approximate i.i.d data samples from the data sources; have some statistics and sampling attribute.

Parameter Estimation

Estimators: Functions that map samples to parameter estimate.

Evaluate:

- Bias: $E[\hat{\theta} - \theta]$.
- Variance: $Var[\hat{\theta}] = E[(\hat{\theta} - E[\hat{\theta}])^2]$.
- Mean Square Error: $MSE(\hat{\theta}) = E[(\hat{\theta} - \theta)^2] = Bias(\hat{\theta})^2 + Var(\hat{\theta})$. If unbiased, $Bias(\hat{\theta}) = 0, MSE(\hat{\theta}) = Var(\hat{\theta})$.

Unbiased Estimator:

- Mean: $\frac{1}{n}\sum_i^n X_i$
- Variance: $\frac{1}{n-1}\sum_i^n (X_i - \bar{X})^2$. (Bessel correction)

Confidence Interval

$X_1, ..., X_n$ i.i.d sampled from a distribution with unknown mean μ and

1. known variance σ^2. From the central limit theorem, we know that $\bar{X} \sim G(\mu, \frac{\sigma^2}{n})$. Therefore, $p = P(\frac{|\bar{X}-\mu|}{\frac{\sigma}{\sqrt{n}}} < z_p) = \Phi(z_p) - \Phi(-z_p) = 2\Phi(z_p) - 1$. After determining the confidence p, we can find the critical value $z_p = \Phi^{-1}(\frac{p+1}{2})$. With z_p, we can say that $\bar{X} \in (\mu - z_p \frac{\sigma}{\sqrt{n}}, \mu + z_p \frac{\sigma}{\sqrt{n}})$ with confidence p, also we can say $\mu \in (\bar{X} - z_p \frac{\sigma}{\sqrt{n}}, \bar{X} + z_p \frac{\sigma}{\sqrt{n}})$ with probability p.
2. unknown variance. We can replace the variance with sample variance with bessel correction. But now $\frac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}}$ does not follow the standard normal distribution, but t distribution with $n-1$ degrees of freedom instead. The other calculations are the same, we simply replace Z distribution with t distribution. $\mu \in (\bar{X} - t_{p,n-1} \frac{S}{\sqrt{n}}, \bar{X} + t_{p,n-1} \frac{S}{\sqrt{n}})$

Hypothesis Test

Given two hypotheses H_0, H_A, we reject H_0 or accept H_A if the samples support H_A; we retain H_0 if the samples do nott support H_A.

- Type-I Error: Declare H_A when H_0 is true, false positive, $P(accept\, H_A | H_0) \leq \alpha$, the significance level.
- Type-II Error: Declare H_0 when H_A is true, false negative, $P(accept\, H_0 | H_A)$

Critical value: x_α for significance level α.

P-value: Probability of the sample or more extreme data are sampled given that H_0 is true. $p = P(X \geq \bar{X} | H_0) \leq \alpha$, accept H_A. (Depends on the H_A, one-sided or two-sided, larger or smaller)