Statistical methods, Formula sheet for final exam

Combinatorics

- Number of ways to choose k out of n objects:

$$\binom{n}{k} = \frac{n(n-1)\cdots(n-k+1)}{k!} = \frac{n!}{(n-k)!k!}$$

Basic probability

Always true:

- $P(A^c) = 1 - P(A)$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- $P(A \cap B) = P(A)P(B|A)$
- $P(A^c) = 1 - P(A)$
- A, B mutually exclusive: $P(A \cup B) = P(A) + P(B)$
- A, B independent: $P(A \cap B) = P(A)P(B)$
- Conditional probability of A given B:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

LTP and Bayes' theorem

$$P(A) = P(A|B)P(B) + P(A|B^c)P(B^c)$$

$$P(B|A) = \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|B^c)P(B^c)}$$

Discrete random variables

- Pmf: $p(x) = P(X = x)$

Special distributions:

- $X \sim \text{bin}(n, p)$: $p(k) = \binom{n}{k} p^k (1-p)^{n-k},\ k = 0, 1, ..., n$ (# of successes)

  $E[X] = np$

- $X \sim \text{geom}(p)$: $p(k) = (1-p)^{k-1} p,\ k = 1, 2, ...$ (wait for first success)

  $E[X] = \dfrac{1}{p}$

Continuous random variables

- Pdf: $f(x) = F^{'}(x), \quad x \in R$
- Cdf: $F(x) = \int_{-\infty}^{x} f(t)dt, \quad x \in R$

Special distributions:

- $X \sim \text{unif}\ [a, b]$: $f(x) = \dfrac{1}{b-a},\ a \leq x \leq b$ (choose "randomly")

  $E[X] = \dfrac{a+b}{2}$

- $X \sim \exp(\lambda)$: $f(x) = \lambda e^{-\lambda x},\ x \geq 0$ (memoryless)

  $E[X] = \dfrac{1}{\lambda}$

- $X \sim N(0, 1)$: $\varphi(x) = \dfrac{1}{\sqrt{2\pi}} e^{-x^2/2},\ x \in R$
- $X \sim N(\mu, \sigma^2) : Z = \dfrac{X - \mu}{\sigma} \sim N(0, 1)$

Expected value

- $E[X] = \sum_k x_x p(x_k)$ if X is discrete with range $\{x_1, x_2, ...\}$
- $E[X] = \int_{-\infty}^{\infty} x f(x) dx$ if X is continuous
- $E[g(X)] = \sum_k g(x_k) p_X(x_k)$
- $E[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x) dx$

Variance

- $\text{Var}[X] = E[(X - \mu)^2] = E[X^2] - (E[X])^2$
- Standard deviation: $\sigma = \sqrt{\text{Var}[X]}$

Sums of random variables

- X and Y independent, a and b constants:

$$E[aX + bY] = aE[X] + bE[Y] \qquad \text{Var}[aX + bY] = a^2\text{Var}[X] + b^2\text{Var}[Y]$$

- $X_1, ..., X_n$ independent random variables with the same distributions, mean μ and varance σ^2, $S_n = X_1 + ... + X_n$.

- $E[S_n] = n\mu \quad$ and $\quad \text{Var}[S_n] = n\sigma^2$

- Central Limit Theorem: S_n is approximately $N(n\mu, n\sigma^2)$ and $\bar{X}$ is approximately $N(\mu, \sigma^2/n)$.

Estimators

- Unbiased: $E[\widehat{\theta}] = \theta$

- We want $\mathrm{Var}[\widehat{\theta}]$ to be as small as possible
- Sample mean $\bar{X}$, unbiased for mean μ, $\mathrm{Var}[\bar{X}] = \sigma^2/n$
- Sample variance $s^2 = \dfrac{1}{n-1}\left(\sum_{k=1}^{n} X_k^2 - n\bar{X}^2\right)$, unbiased for variance σ^2
- Standard error: $\sqrt{\mathrm{Var}[\widehat{\theta}]}$

Confidence intervals

- For μ in $N(\mu, \sigma^2)$ where σ^2 is known:

$$\mu = \bar{X} \pm z\frac{\sigma}{\sqrt{n}} \quad (q)$$

Use standard normal distribution, $\Phi(z) = (1+q)/2$.

- For μ in $N(\mu, \sigma^2)$ where σ^2 is unknown:

$$\mu = \bar{X} \pm t\frac{s}{\sqrt{n}} \quad (q)$$

Use t distribution, $\alpha = 1 - (1+q)/2$, $\nu = n - 1$.

- For unknown proportion p:

$$p = \widehat{p} \pm z\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} \quad (q)$$

- For two unknown proportions p_1 and p_2:

$$p_1 - p_2 = \widehat{p}_1 - \widehat{p}_2 \pm z\sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}} \quad (q)$$

In both cases, z is such that $\Phi(z) = (1+q)/2$.

Estimation methods

1. The maximum likliehood estimator (MLE) $\widehat{\theta}$ maximizes the likelihood function:

$$L(\theta) = \prod_{k=1}^{n} f_\theta(X_k)$$

To find maximum, (i) take logarithm, (ii) differentiate w.r.t. θ and set $= 0$.

2. The rth moment and rth sample moment are:

$$\mu_r = E[X^r] \quad \text{and} \quad \widehat{\mu}_r = \frac{1}{n} X^r$$

The method of moments estimator (MOME) expresses the parameter as a function of moments, $\theta = g(\mu_1, ..., \mu_r)$ and estimates it with the same funtion of the sample moments, $\widehat{\theta} = g(\widehat{\mu}_1, ..., \widehat{\mu}_r)$. Start with the first moment, if it is not enough, go on to the second, and so on.

Linear regression

- Model: $Y = a + bx + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2)$
- Observations: $(x_1, Y_1), ..., (x_n, Y_n)$ where $Y_k \sim N(a + bx_k, \sigma^2)$
- Notation:

$$\begin{aligned} S_x &= \sum_{k=1}^{n} x_k, \quad S_Y = \sum_{k=1}^{n} Y_k \\ S_{xx} &= \sum_{k=1}^{n} x_k^2, \quad S_{xY} = \sum_{k=1}^{n} x_k Y_k \end{aligned}$$

- Estimators:

$$\widehat{b} = \frac{nS_{xY} - S_x S_Y}{nS_{xx} - S_x^2}$$

$$\widehat{a} = \bar{Y} - \widehat{b}\bar{x}$$

Estimated regression line: $y = \widehat{a} + \widehat{b}x$