Summary Midterm

February 24, 2023 11:27

G1.1.1 Random Sample

A sample $X_1, \dots, X_n$ from a population with CDF F_X:

- $X_1, \dots, X_n$ are independent
- $X_1, \dots, X_n$ have the same distribution, same CDF F_X

A function of a sample, such as $h(X_1, \dots, X_n)$, is a statistic

G1.1.2 Mean and Standard Deviation

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

G1.1.3 Linear Combination of RVs

For $Y = a_1X_1 + \cdots + a_nX_n$, we have:
$E[Y] = a_1E[X_1] + \cdots + a_nE[X_n]$

If $X_1, \dots, X_n$ are independent, then:
$V[Y] = a_1^2V[X_1] + \cdots + a_n^2V[X_n]$

+ Theorems

- For $X_1, \dots, X_n$ independent such that $X_i \sim N(\mu_i, \sigma_i^2)$ and $Y = a_1X_1 + \cdots + a_nX_n$, then:
$$Y \sim N\left(\sum_{i=1}^{n} a_i\mu_i\,,\ \sum_{i=1}^{n} a_i^2\sigma_1^2\right)$$
- If $X_1, \dots, X_n$ is a sample with distribution $N(\mu, \sigma^2)$, then:
$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

G1.1.5 T-Distribution

For $Z \sim N(0,1)$ and $U \sim \chi^2(r)$ that are independent:
$$T = \frac{Z}{\sqrt{\left(\frac{U}{r}\right)}} \sim t(r)$$

+ Quantiles of the T-Distribution

$P(T > t_\alpha(r)) = \alpha$ with $t_\alpha(r)$, the upper quantile of order α

Properties:

- $t_{1-\alpha}(r) = -t_\alpha(r)$ since it's symmetric about $t = 0$
- As $r \to \infty$, the t distribution goes to $N(0,1)$

G1.1.4 Standard Normal Distribution

For $X \sim N(\mu, \sigma^2)$, the standard normal distribution is:
$$Z = \frac{X - \mu}{\sigma} \sim N(0,1)$$

For a sample from a population with distribution $N(\mu, \sigma^2)$:
$$\frac{\bar{X} - \mu}{\left(\frac{\sigma}{\sqrt{n}}\right)} \sim N(0,1), \quad \text{since } \bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

+ Theorems

- If $Z \sim N(0,1)$, then $Z^2 \sim \chi^2(r)$
- For $X_1, \dots, X_n$ independent RVs such that $X_i \sim \chi^2(r_i)$:
If $W = X_1 + \cdots + X_n$, then $W \sim \chi^2(r_1 + \cdots + r_n)$
- For $X_1, \dots, X_n$ a sample from a $N(\mu, \sigma^2)$ distribution:
If $W = \sum_{i=1}^{n}\left(\frac{X_i - \mu}{\sigma}\right)^2$, then $W \sim \chi^2(n)$
- For $X_1, \dots, X_n$ a sample from a $N(\mu, \sigma^2)$ distribution:
  - $\bar{X}$ and S^2 are independent
  - $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$

G2.1.3 Binomial Approximation with a Normal

Define the following:

- $Y \sim binom(n,p)$ with CDF $F_Y(y)$
- $W = \frac{Y - np}{\sqrt{np(1-p)}}$, the standardized version of Y

For n where $np \geq 5$ and $np(1-p) \geq 5$, apply the theorem:
$F_Y(y) \approx \phi\left(\frac{y + 0.5 - np}{\sqrt{np(1-p)}}\right)$ where ϕ is the CDF of $N(0,1)$

G1.1.6 F-Distribution

For $U_1 \sim \chi^2(r_1)$ and $U_2 \sim \chi^2(r_2)$: $F = \frac{(U_1/r_1)}{(U_2/r_2)} \sim F(r_1, r_2)$

+ Quantiles of the F-Distribution

$P\left(F > F_\alpha(r_1, r_2)\right) = \alpha$ with $F_\alpha(r_1, r_2)$, the upper quantile of order α

Properties:

- $F \sim F(r_1, r_2) \to \frac{1}{F} \sim F(r_2, r_1)$
- In another form: $F_{1-\alpha}(r_1, r_2) = \frac{1}{F_\alpha(r_2, r_1)}$

G2.1.1 Order Statistics

Given $X_1, \dots, X_n$, a sample with CDF $F(x)$ and PDF $f(x)$:
Sorting gives the order statistics $Y_1, \dots, Y_n$ where $Y_1 \leq \dots \leq Y_n$

CDF and PDF of Y_r for $1 \leq r \leq n$:

- $F_r(y) = \sum_{k=r}^{n} \binom{n}{k} \left(F(y)\right)^k \left(1 - F(y)\right)^{n-k}$
- $f_r(y) = \binom{n}{r-1, 1, n-r} \left(F(y)\right)^{r-1} \left(f(y)\right) \left(1 - F(y)\right)^{n-r}$

Binomial Formulas:

- $\binom{n}{k} = \frac{n!}{k!\,(n-k)!}$
- $\binom{n}{r-1,\ 5,\ n-r} = \frac{n!}{(r-1)!\,1!\,(n-r)!}$

G1.2.1 Sample Percentile

For $0 < p < 1$, the $(100p)$th sample percentile denoted $\tilde{\pi}_p$ is:

- If $(n+1)p$ is an integer, then $\tilde{\pi}_p = y_{(n+1)p}$
- Otherwise, then: $\exists r, a, b$ such that $(n+1)p = r + \frac{a}{b}$

  So $\tilde{\pi}_p = \left(1 - \frac{a}{b}\right) y_r + \left(\frac{a}{b}\right) y_{r+1}$

G1.2.2 Quartiles

- $q_1 = \tilde{\pi}_{0.25}$
- $q_2 = \tilde{m} = \tilde{\pi}_{0.5}$, the median
- $q_3 = \tilde{\pi}_{0.75}$
- $IQR = q_3 - q_1$

G2.1.2 Population Percentile

The $(100p)$th population percentile, denoted π_p satsifies:
$P[X \leq \pi_p] = p$

Given the order statistics $Y_1, \dots, Y_n$, and the CDF of a binomial distribution $F(x)$:
$P[Y_i < \pi_p < Y_j] = F(j-1) - F(i-1)$

G1.2.3 Central Tendencies

- Sample Mean: $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$
- Sample Median: $\tilde{m} = \begin{cases} y_{(n+1) \times \frac{1}{2}}, & n \text{ is odd} \\ \dfrac{y_{\frac{n}{2}} + y_{\left(\frac{n}{2}\right)+1}}{2}, & n \text{ is even} \end{cases}$

+ Theorems

- The sample mean $\bar{x}$ minimizes : $\sum_{i=1}^{n} (x_i - \theta)^2$
- The median of the sample $\tilde{m}$ minimizes: $\sum_{i=1}^{n} |x_i - \theta|$

G4.1.2 Measures of Dispersion

- Sample standard deviation: $S = \sqrt{S^2}$
- Sample Variance S^2
- Range: $Y_n - Y_1$
- $IQR = q_3 - q_1$

G1.3.1 Point Estimation

- Sample standard deviation: $S = \sqrt{S^2}$
- kth sample moment: $M_k = \frac{1}{n} \sum_{i=1}^{n} X_i^k$
- kth sample central moment: $\frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^k$

+ Expected Value Formulas

- $E[X] = \int_{-\infty}^{\infty} a f_X(a) da$
- $E[X^2] = \int_{-\infty}^{\infty} a^2 f_X(a)\, da$
- $E[X^k] = M_k$
- $E[X^2]$, denoted $V = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2$

G1.3.2 Statistics

- If a statistic $h(X_1, \dots, X_n)$ estimates θ, then it's an estimator for θ, denoted by $\hat{\theta} = h(X_1, \dots, X_n)$
- A point estimate $\hat{\theta}$ for θ is $h(x_1, \dots, x_n)$ where $x_1, \dots, x_n$ are observed values from a sample
- An estimator $\hat{\theta}$ for θ is unbiased if $E[\hat{\theta}] = \theta$

G1.3.3 Maximum Likelihood Estimator (MLE)

We want to maximize the likelihood functions:

- $L(\theta) = \prod_{i=1}^{n} f(x_i, \theta)$
- $\ln(L(\theta)) = \sum_{i=1}^{n} \ln\left(f(x_i, \theta)\right)$

We maximize the likelihood functions using either one:

- $\frac{\delta L(\theta)}{\delta \theta} = 0$
- $\frac{\delta \ln(L(\theta))}{\delta \theta} = 0$

Maximized at $(\theta_1, \dots, \theta_k) = \left(h_1(x_1, \dots, x_n), \dots, h_k(x_1, \dots, x_n)\right)$:

- The MLEs are $\hat{\theta}_i = h_i(X_1, \dots, X_n)$
- The maximum likelihood estimates are $\hat{\theta}_i = h_i(x_1, \dots, x_n)$

G1.3.4 Method of Moments

Given a random sample $X_1, \dots, X_k$ from a population with unknown θ_i, we need to estimate the unknown parameters

We get $E[X] = M_1$, if we can solve for θ_i, we stop otherwise:
We get $E[X^2] = M_2$, we get a system of equations, and if we can solve for θ_i, we stop, otherwise, we get $E[X^3] = M_3$ and so on…

G5.1.1 Linear Regression

Study the relationship between the independent variable x, the regressor, and the dependent variable Y, the response variable

Let $(x_1, Y_1), \ldots, (x_n, Y_n)$ be the sample for the regression model

G5.1.2 Linear Regression Assumptions

- For the general regression model:
  - $E[Y] = \alpha_1 + \beta x$
  - $Y = \alpha_1 + \beta x + \epsilon$ where $\epsilon \sim N(0, \sigma^2)$ so $E[\epsilon] = 0$

- For the random sample:

$$\left.\begin{aligned} &\alpha_1 = \alpha - \beta\bar{x} \\ &Y_i = \alpha + \beta(x_i - \bar{x}) + \epsilon_i \\ &\epsilon_i \sim (iid)\ N(0, \sigma^2) \\ &E[Y_i] = \alpha + \beta(x_i - \bar{x}) \\ &Var[Y_i] = \sigma^2 \end{aligned}\right\} Y_i \sim N(\alpha + \beta(x_i - \bar{x}), \sigma^2)$$

G5.1.3 Notation for Linear Regression

- $S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \left(\sum_{i=1}^{n} x_i^2\right) - n\bar{x}^2$

- $S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \left(\sum_{i=1}^{n} y_i^2\right) - n\bar{y}^2$

- $S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \left(\sum_{i=1}^{n} y_i^2\right) - n\bar{x}\bar{y}$

G5.1.4 MLEs for $\alpha,\ \beta, \sigma^2$

- $\hat{\alpha} = \bar{y}$
- $\hat{\beta} = \frac{S_{xy}}{S_{xx}}$
- $\hat{\sigma}^2 = \frac{S_{yy} - \hat{\beta}S_{xy}}{n}$

+ Common Formulas

+ CDF $F_X(a)$ Formulas

- $F_X(a) = P[X \leq a] = \int_{-\infty}^{a} f_X(a)\, da$
- $P[a < X \leq b] = F_X(b) - F_X(a) = \int_{a}^{b} f_X(a)\, da$

+ Bernoulli Distribution $X \sim bern(p)$

- $P_X(a) = p^a(1-p)^{1-a}$
- $E[X] = p$
- $\hat{p} = \frac{1}{n}\sum^{i=1} X_i = \bar{X}$
- $Var[X] = p(1-p)$

G4.2.1 QQ-Plots Normality

If the theoretical data line and actual data points are close, then the QQ-Plot represents a normally distributed population

+ Binomial Distribution $X \sim binom(n,p)$

- $P_X(a) = \binom{n}{a} p^a(1-p)^{n-a}$
- $E[X] = np$
- $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}$
- $Var[X] = np(1-p)$

+ Geometric Distribution $X \sim geom(p)$

- $P_X(a) = p(1-p)^{a-1}$
- $E[X] = \frac{1}{p}$
- $\hat{p} = \frac{n}{\sum_{i=1}^{n} X_i}$
- $Var[X] = \frac{1-p}{p^2}$

+ Poisson Distribution $X \sim pois(\lambda)$

- $P_X(a) = \frac{\lambda^a e^{-\lambda}}{a!}$
- $E[X] = \lambda$
- $\hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}$
- $\mathrm{Var}[X] = \lambda$

+ Exponential Distribution $X \sim \exp(\lambda)$

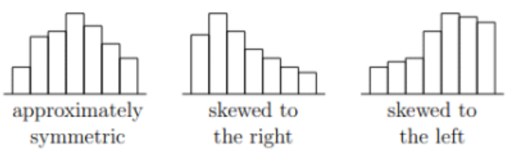
- $f_X(a) = \lambda e^{-\lambda a}$
- $E[X] = \frac{1}{\lambda}$
- $F_X(a) = 1 - e^{-\lambda a}$
- $Var[X] = \frac{1}{\lambda^2}$
- $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X} \rightarrow \hat{\lambda} = \frac{1}{\hat{\theta}} = \frac{1}{\bar{X}}$

+ Normal Distribution $X \sim N(\mu, \sigma^2)$

- $f_X(a) = \frac{1}{\sigma\sqrt{2\pi}} \times e^{\frac{-(x-\mu)^2}{2\sigma^2}}$
- $E[X] = \mu$
- $F_X(a) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{a} e^{\frac{y^2}{2}} dy$
- $Var[X] = \sigma^2$
- $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}$
- $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$

G4.1.1 Descriptive Statistics

Histogram:



Boxplot: