Statistical Concepts Review Notes

Krzysztof Zaremba

February 16, 2024

Qualitative Methods of Forecasting

- Delphi Method - Multiple experts asked about an issue through questionnaires. Answers summarized and sent back to experts for feedback. Iterated until consensus.
- Panel of Experts - Experts gathered together to discuss the issue
- Brainstorming - Free generation of ideas, not need experts
- Focus Group - Representative group of people (not experts) providing insights about a perticular issue

Types of data and variables

- Types of data
  - Primary: collected by researchers for their specific question
  - Secondary: reused by researchers, collected by someone else for a different purpose
- Types of datasets
  - Cross Section: Multiple units, each observed only at a one point in time
  - Longitudinal: One unit, observed over multiple time periods
  - Panel: Multiple units, each observed at multiple time periods
- Types of variables
  - Categorical - no numerical meaning, can't add/divide/multiply values
    * Nominal: divide observations into groups, no ordering of groups
    * Ordinal: divide observations into groups, groups can be ordered
  - Numerical - can add/divide/multiply values

    * Discrete: Finite number of values, countable quantities
    * Continuous: Values within a given range, can be infinitely divided

Graphical summaries of variables

- Categorical variables
  - Frequency Tables
  - Bar Charts
  - Pie Charts
  - Treemaps
- Numerical variables
  - Dotplot
  - Frequency Distribution
  - Histograms
  - Box and Whiskers plot

Parameters and statistics

- Population: The entire population we are interested in
- Sample: A (randomly) chosen group of units from the population
- Parameter: The number or property that characterizes the population that we want to know
- Statistics: A guess of the parameter calculated from the sample

Measures of Central Tendency and Dispersion and Association

- **Mean**
  - Sample mean: $\bar{x} = \frac{\sum_i x_i}{n}$
  - Population mean $\mu = E(x_i) = \frac{\sum_i x_i}{N}$ (if discrete)
  - Expectation properties:
    * If a and b are constants, then $E(ax_i + b) = aE(x_i) + b$
    * $E(x_i + x_j) = E(x_i) + E(x_j)$
- **Median**: The middle value when the data is ordered.

- **Mode**: The most frequently occurring value in the data set.
- **Percentiles** The value below which a given percentage of observations falls. If the data is ordered from smallest to largest, p percentile K_p corresponds to observation number $i = \frac{p(n+1)}{100} \to K_p = x_(i)$ In other terms: $P(X < K_p) = p$
- **Variance**:
  - Sample variance: $s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1} = \frac{\sum x_i^2 - n\bar{x}^2}{n-1}$
  - Population variance $\sigma^2 = var(x_i) = E[(x_i - \mu)^2] = E(x_i^2) - [E(x_i)]^2$
  - Variance properties:
    * If a and b are constants, then $var(ax_i + b) = a^2 var(x_i)$
    * $var(x_i + x_j) = var(x_i) + var(x_j) + 2cov(x_i, x_j)$
- **Standard Deviation** (s): The square root of the variance. $s = \sqrt{s^2}$.
- **Range**: The difference between the largest and smallest value.
- **Inter-quartile range**: The difference between the third and first quartile
- **Coefficient of Variation** (CV): A standardized, measure of dispersion which does not depends on the units $CV = \frac{\sigma}{|\mu|}$. Can use to compare across variables
- **Covariance**: A measure of the joint variability of two random variables. For sample covariance.
  - Sample covariance: $\hat{\sigma_{xy}} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n-1} = \frac{\sum_i x_i y_i - n\bar{x}y}{n-1}$
  - Population covariance $\sigma_{xy} = cov(x, y) = E[(x_i - \mu_x)(y_i - \mu_y)]$
  - Covariance properties:
    * If a and b are constants, $cov(ax_i, by_i) = abcov(x, y)$
    * If c and d are constants, $cov(x_i + c, y_i + d) = cov(x, y)$
- **Correlation Coefficient** (r): Standardized form of covariance, giving values between -1 and 1.

  $$r = \frac{\text{cov}(X, Y)}{s_X s_Y}$$

  where s_X and s_Y are the standard deviations of X and Y, respectively. Can use to compare across variables
- Contingency table: measure of association between categorical variables based on conditional probabilities (share within subgroups). Shows exhaustive and mutually exclusive list of categories

Probability Distributions and Statistical Inference

- **Cumulative Distribution Function** (CDF): $F(x) = P(X \leq x)$.
- **Probability Density Function** (PDF) for continuous variables: $f(x)$ where $P(a \leq X \leq b) = \int_a^b f(x)\,dx$.
- Properties of the normal:
  - If $x_i \sim N(\mu, \sigma)$, then $\frac{x_i - \mu}{\sigma} \sim N(0, 1)$
  - If $x_i \sim N(\mu, \sigma)$, a and b are constant, then $ax_i + b \sim N(a\mu + b, |a|\sigma)$
  - If $x_i \sim N(\mu_x, \sigma_x)$, and $y_i \sim N(\mu_y, \sigma_y)$, then $x_i + y_i \sim N(\mu_x + \mu_y, \sqrt{\sigma_x^2 + \sigma_y^2})$
- Random sample: a number of iid observations drawn at random. iid- independently and identically distributed
  - independent: draw of one observation does not change the probability of a draw of the next observation
  - identically distributed: come from the same distribution, have same mean and variance
- Sampling distribution: distribution of a sample statistic/estimator - of a function of random sample (eg sum, mean etc)
- **Central Limit Theorem**:
  - The CLT applies to the distribution of the sample mean ($\bar{X}$), sums ($\sum X_i$), and standardized means ($\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$). They converge to normal distributions under the assumptions below.
  - **Assumptions**: The samples should be independent and identically distributed (i.i.d.). The theorem holds exactly in the limit as sample size n tends to infinity, and approximately for finite sample sizes large enough (usually $n \geq 30$ is considered sufficient).

Estimators and Their Properties

Estimator: A statistic used to infer the value of an unknown parameter in a statistical model. **Bias of an Estimator**: The difference between the expected value of the estimator and the true value of the parameter.

$$\text{Bias}(\hat{\theta}) = E(\hat{\theta}) - \theta$$

Variance of an Estimator: The variability of the estimator.

$$\text{Var}(\hat{\theta}) = E\left[(\hat{\theta} - E(\hat{\theta}))^2\right]$$

Mean Squared Error (MSE): The average of the squares of the errors—that is, the average squared difference between the estimated values and the actual value.

$$\text{MSE}(\hat{\theta}) = E\left[(\hat{\theta} - \theta)^2\right] = \text{Bias}(\hat{\theta})^2 + \text{Var}(\hat{\theta})$$

Example: Sample Mean as an Estimator: The sample mean $\bar{X}$ is an unbiased estimator of the population mean μ.

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

- – $Bias(\bar{x}) = E(\bar{x}) - \mu = 0$
  – $MSE(\bar{x}) = var(\bar{x}) = \frac{var(x)}{n}$

Central Limit Theorem (CLT)

- Let X_i be a random variable from any distribution. Then the sample mean ($\bar{X}$), sums ($\sum X_i$), and standardized means ($\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$) converge to normal distribution.
- **Assumptions**: The samples should be independent and identically distributed (i.i.d.), with a finite mean μ and finite variance σ^2. The theorem holds exactly in the limit as sample size n tends to infinity, and approximately for finite sample sizes large enough (usually $n \geq 30$ is considered sufficient).

Confidence Intervals

- **Confidence Interval for Mean**: What distribution to use?
  - Normal distribution $\bar{x} \pm z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$ if:
    * If $n > 30$, known variance
    * If $n > 40$, unknown variance
    * If $n < 30$, known variance and $x \sim N(\mu, \sigma)$
  - Student t with n-1 degrees of freedom $\bar{x} \pm t_{\frac{\alpha}{2},n-1} \frac{s}{\sqrt{n}}$ if
    * If $n < 40$, unknown variance and $x \sim N(\mu, \sigma)$
  - All other cases (small n and unknown distibution of x), can't do anything
- **Confidence Interval for Variance**: Based on the chi-square distribution, the confidence interval for the population variance σ^2 is given by:

$$\left(\frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2},n-1}}, \frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2},n-1}}\right)$$

where $\chi^2_{\frac{\alpha}{2},n-1}$ and $\chi^2_{1-\frac{\alpha}{2},n-1}$ are the chi-square critical values at the desired confidence level.

- **Interpretation of Confidence Intervals**: We are $1-\alpha\%$ confident that the true value of the parameter is within this range.