Stat 525 Regression Analysis

Lecture 0 : Review of statistical concepts

Zheni Utic
Department of Mathematics and Statistics, University of Massachusetts Amherst

Statistics : population and sample

- Statistics

- make generalizations (inference) about the characteristics of a population based on information contained in a sample from that population.

  i) selection of the sample from the population of our interest

  ii) statistical inference about the characteristics of the population

- Selection of the sample from the population of our interest

- The population is the collection of all elements (individuals, items, or objects) whose characteristics are being studied.

  Our interest is to study characteristics of (the units of) the population. (Unknown) quantifiable properties of a characteristic of interest (e.g., an average or a proportion) are called parameters (denoted by θ)

- Due to cost and time considerations, we typically obtain a sample, a portion of the population selected for study (a representative sample).

- A variable is a characteristic under study that is (quantitative/qualitative) quantity and assume different values for different elements.

  A random variable (denoted by a capital letter such as X, or Y) is the *a priori* unknown value of the variable of an element randomly sampled from a population.

- We randomly take a sample of n elements from the population and record the variable of each sampled element. Henceforth, the word random sample consists of unknown *a priori* numbers (so random variables).

  $X_1, X_2, \ldots, X_n$ represents a random sample of size n and $x_1, x_2, \ldots, x_n$ are observed values of $X_1, X_2, \ldots, X_n$ when the study is carried out.

- Usually assume that $X_1, X_2, \ldots, X_n$ is a random sample from a population with probability distribution $f(x; \theta)$

Mercury study

Suppose we are interested in estimating the **average/mean** Mercury concentration μ in a lake. We decide to obtain 50 water samples drawn from a lake (by using a suitable sampling method)

- population : all water samples that can be taken from a lake

- sample : a subset of all water samples

- parameter : mean Mercury concentration μ in the lake

- random variable (denoted by X): Mercury concentration (quantitative)

- $X_1, \ldots, X_{50}$: a random sample of size 50 from the population

Statistics : statistical inference and its tools

- Statistical inference about the unknown parameters using $X_1, \ldots, X_n$
  - generalizes the information contained in the sample to the population,

    assess how closely sample characteristics resemble population characteristics,

    assess the likelihood of making wrong decisions regarding the true value of population parameters

- Tools for statistical inference
  - point and interval estimation
  - hypothesis testing
  - prediction
  - more..

Statistical inference - point estimation

Suppose $X_1, \ldots, X_n$ be a random sample from a population with θ or $f(x; \theta)$.

- A statistic is a function of $X_1, \ldots, X_n$ and known constants (i..e, random variable) and has a probability distribution under repeated sampling of the population (called sampling distribution). The sampling distribution of a statistic is the population of all possible values for that statistic.

  (e.g.) assume we repeatedly take samples of a given size from the population and calculate the sample mean, $\bar{x}$ for each sample. Different samples will lead to different sample means. The distribution of these means is the "sampling distribution of $\bar{X}$" (for the given sample size).

- Point estimator for θ

  - a statistic is used as an estimator $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$ (a formula) for θ and $\hat{\theta}$ has its sampling distribution

    an estimate $\hat{\theta}(x_1, \ldots, x_n)$ is a specific value of the estimator at $x_1, \ldots, x_n$

    a standard error for $\hat{\theta}$, $dev(\hat{\theta})$, is a square root of variance of $\hat{\theta}$, $\sqrt{dev^2(\hat{\theta})}$

(e.g.) suppose $X_1, \ldots, X_n$ be a random sample from a population with mean $\mu = E(X_i)$ and variance $\sigma^2 = Var(X_i)$.

i) sample mean $\hat{\mu} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ for μ, and its standard error is $\sqrt{\sigma^2/n}$

ii) sample variance $\hat{\sigma}^2 = s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ for σ^2

iii) sample standard deviation $s = \sqrt{s^2}$ for σ

(e.g.) suppose $X_1, \ldots, X_n$ be a random sample from a population with $N(\mu, \sigma^2)$.

What is the sampling distribution of $\hat{\mu} = \bar{X}$?

- Properties of point estimator $\hat{\theta}$
  - MSE(Mean Square Error) of $\hat{\theta} = E\left[(\hat{\theta} - \theta)^2\right] = \sigma^2(\hat{\theta}) + Bias(\hat{\theta})^2$ where $Bias(\hat{\theta}) = E(\hat{\theta}) - \theta$
  - unbiasedness and minimum variance

- Estimation method
  - Method of Maximum Likelihood

    (e.g.) suppose $X_1, \ldots, X_n$ be a random sample from a population with $N(\mu, \sigma^2)$ where σ^2 is known we can for example find an estimator for μ using Method of Maximum Likelihood.
  - Method of Least Squares

    (e.g.) suppose $X_1, \ldots, X_n$ be a random sample from a population with mean $\mu = E(X_i)$. We can find an estimator for μ using Method of Least Squares.

Statistical inference - interval estimation

Suppose $X_1, \ldots, X_n$ be a random sample from a population with θ.

- $100(1-\alpha)\%$ (two-sided) confidence interval for θ, $[\hat{\theta}_L, \hat{\theta}_U]$
  - $\hat{\theta}_L = \hat{\theta}_L(X_1, \ldots, X_n), \hat{\theta}_U = \hat{\theta}_U(X_1, \ldots, X_n)$
  - a random interval of enclosing θ with a probability $(1-\alpha)$ under repeated sampling :

    $1-\alpha$ (confidence coefficient) = $P(\hat{\theta}_L \leq \theta \leq \hat{\theta}_U)$= fraction of the time, in repeated sampling, that $[\hat{\theta}_L, \hat{\theta}_U]$ will contain θ.
- Large-sample confidence intervals (by using Central Limit Theorem)
  - useful when n is large and there is no distribution assumption on $X_1, \ldots, X_n$
  - Approximate probability distribution of $\frac{\hat{\theta}-\theta}{\sigma_{\hat{\theta}}}$ is a standard normal distribution, $N(0,1)$ as long as n is large (i.e., $\frac{\hat{\theta}-\theta}{\sigma_{\hat{\theta}}} \sim_{approx} N(0,1)$)

    Then a (approximate) $100(1-\alpha)\%$ two-sided confidence interval for θ is $[\hat{\theta}_L, \hat{\theta}_U] = [\hat{\theta} - z_{\alpha/2}\sigma_{\hat{\theta}}, \hat{\theta} + z_{\alpha/2}\sigma_{\hat{\theta}}]$

- Confidence intervals under the **normal** populations

  1) Suppose $X_1, \ldots, X_n$ be a random sample from $N(\mu, \sigma^2)$.

  i) $100(1-\alpha)\%$ confidence intervals for μ by using $\frac{\bar{X}-\mu}{s/\sqrt{n}} \sim t(n-1)$

  : $[\, \bar{X} - t(1-\alpha/2; n-1)s/\sqrt{n}, \bar{X} + t(1-\alpha/2; n-1)s/\sqrt{n}\,]$

  ii) $100(1-\alpha)\%$ confidence intervals for σ^2 by using $\frac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1)$

  : $\left[\frac{(n-1)s^2}{\chi^2(1-\alpha/2;n-1)}, \frac{(n-1)s^2}{\chi^2(\alpha/2;n-1)}\right]$

  2) Suppose $X_1, \ldots, X_{n_1}$ be a random sample from $N(\mu_1, \sigma^2)$ and $Y_1, \ldots, Y_{n_2}$ be a random sample from $N(\mu_2, \sigma^2)$.

  i) $100(1-\alpha)\%$ confidence intervals for $\mu_1 - \mu_2$ by using $\frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{\sqrt{s_p^2(\frac{1}{n_1}+\frac{1}{n_2})}} \sim t(n_1+n_1-2)$ where $s_p^2 = \frac{\sum_{i=1}^{n_1}(X_i-\bar{X})^2+\sum_{i=1}^{n_2}(Y_i-\bar{Y})^2}{n_1+n_2-2}$

  : $\left[(\bar{X}-\bar{Y}) - t(1-\alpha/2; n_1+n_2-2)\sqrt{s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}, (\bar{X}-\bar{Y}) + t(1-\alpha/2; n_1+n_2-2)\sqrt{s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}\right]$

3) Suppose $X_1, \ldots, X_{n_1}$ be a random sample from $N(\mu_1, \sigma_1^2)$ and $Y_1, \ldots, Y_{n_2}$ be a random sample from $N(\mu_2, \sigma_2^2)$.

i) $100(1-\alpha)\%$ confidence intervals for $\frac{\sigma_1^2}{\sigma_2^2}$ by using $\frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} \sim F(n_1-1, n_2-1)$ where $s_1^2 = \frac{\sum_i^{n_1}(X_i-\bar{X})^2}{n_1-1}$ and $s_2^2 = \frac{\sum_i^{n_2}(Y_i-\bar{Y})^2}{n_2-1}$

: $\left[\frac{s_1^2}{s_2^2}\frac{1}{F(1-\alpha/2;n_1-1,n_2-1)}, \frac{s_1^2}{s_2^2}\frac{1}{F(\alpha/2;n_1-1,n_2-1)}\right]$

Statistical inference - hypothesis testing

Suppose $X_1, \ldots, X_n$ be a random sample from a population with $f(x; \theta)$.

- Elements of a statistical test

  1) Hypothesis

  : null hypothesis, H_0 (what we would like to refute). e.g., $H_0 : \theta = \theta_0$

  : alternative hypothesis H_a (what we would like to support by evidence in the sample). e.g., $H_a : \theta \neq \theta_0$ (two-sided) or $H_a : \theta > \theta_0$ (one-sided)

  2) Test statistic, $U = U(X_1, \ldots, X_n)$

  : a function of the random sample whose distribution under H_0 (null distribution) is known and can thus be used as reference

  : we can reject H_0 in favor of H_a if the observed value of U, $u_{obs} \equiv u(x_1, \ldots, x_n)$ is very extreme with respect to what one would expect under the null distribution

3) Probability of making mistake

: α=P(Type I error)=P(false positive)=P(reject H_0 | H_0 is true)

(level or significance level)

: β=P(Type II error)=P(false negative)=P(not reject H_0 | H_a is true)

: power of the test = 1-β (so power function = a function expressing power for each value in H_a)

: α and β are in trade-off; test statistics are evaluated based on their power function, once α is fixed

4) Decision rule

Suppose α is fixed.

i) use Rejection Region(RR) associated with U and α, RR_α

: RR_α specifies the values of U for which H_0 is to be rejected in favor of H_a. So the rule is as follows : first, compute the value of the test statistic for an observed sample, u_{obs}. If u_{obs} falls in the RR, reject H_0.

ii) use the p-value

: the p-value associated with u_{obs} is the probability that, under H_0, U would take the observed value u_{obs}, or a value even more extreme in the direction defined by the H_a.

(e.g.) if the null distribution of U is symmetric and H_a is two-sided, the p-value is

: the smaller the p-vaule, the stronger the evidence against H_0

: reject H_0 if the computed p-value $\leq \alpha$

- Review Appendix A.6-A.9 in KNN and solve self-test questions

Statistical modeling

The structural form of the model describes the patterns of interactions and associations between the variables of interest. The model parameters provide measures of strength of associations. In models, the focus is on inference on the model parameters. The basic inference tools (e.g., point estimation, confidence intervals and hypothesis testing) will be applied to these parameters.

- Objective
- Model structure (e.g. variables, formula, equation)
- Model assumptions
- Inference on model parameter and interpretation
- Model fit (e.g. goodness-of-fit)
- Model selection