Statistical Inference

F. Song

June 2018

1 Hypothesis testing

Hypothesis testing is concerned with making decisions with data. The null hypothesis is assumed true and statistical evidence is required to reject it in favor of a research or alternative hypothesis .

Every time we perform a hypothesis test, we will follow: (1) we'll make an initial assumption about the population parameter (2) we'll collect evidence or else use somebody else's evidence

(3) Based on the available evidence, we'll decide whether to "reject" or "not reject" our initial assumption. To make the decision, there are two approaches:

1. one is called the "critical value" approach:define a threshold value, called a "critical value" such that if our "test statistic" is more extreme than the critical value, we reject the null hypothesis
   (a) state H_0 and H_A
   (b) calculate the test statistic
   (c) Determine the critical region
   (d) Make a decision. Determine if the test statistic falls in the critical/rejection region.If it does, reject the null hypothesis. Otherwise, accept the null hypothesis. if we reject the null hypothesis when the null hypothesis is in fact true, we say we've committed a type 1 error. we wanted to minimize our chance of making a Type 1 error! In general, we denote $\alpha = P(Type1error) =$ "$significancelevelof test$". We want to minimize α, so typically α are 0.01,0.05,0.10. The quantileprobability associated with the test statistics is called the **P value**. If p value $< \alpha$(usually, $\alpha = 0.05$), reject the null hypothesis. The P-value is the probability that we'd observe a more extreme statistic than we did if the null hypothesis were true
2. and the other is called the "p-value " approach

2 Confident level and Hypothesis test

t distribution has thicker tails than the normal distribution. It's indexed by a degree of freedom and it gets more like a standard normal as the degree of freedom get larger. It assumes that the underlying data set are iid Gaussian with the result that $\frac{\bar{X}-\mu}{S/\sqrt{n}}$ follows t distribution with $n-1$ degrees of freedom.

The standard error of an estimator reflects how it varies under repeated sampling. These standard errors can be used to compute confidence intervals. A 95% confidence interval is defined as a range of values such that with 95% probability, the range will contain the true unknown value of the parameter.

Standard error can also be used to perform hypothesis tests on the parameters. Hypothesis test is a test of a certain value of a parameter. To test the null hypothesis, we compute a t-statistic, given by

$$t = \frac{\hat{(}\beta_1) - 0}{SE(\hat{\beta}_1)}$$

Using software, it is easy to computer the *p-value.* which is the probability of observation any value equal to $|t|orlarger$.

t statistics of bigger than about 2 is significant at p value of 0.5

t-statistis, p value R^2,F-statistics

In statisticas, a hypothesis test calculates some quantity under a given assumption. The result of the rest allow us to interpret whether the assumption holds or whether the assumption has been violated. The assumption of a statistical test is called the **null hypothesis, or H_0, or default assumption**, a violation of the test's assumption is called the first hypothesis, or H_1. A statistical hypothesis test may return a value called p -value. This is a quantity that we can use to interpret or quantify the results of the test and either reject of fail to reject the null hypothesis. This is done by comparing the p-value to a thershold value chosen beforehand called the significance level α. A common value used for α is 0.05. A smaller α value suggests a more robust interpretation of the null hypothesis. The p value is compared to the prechosen α value.

If p value $>$ α: Fail to reject the null hypothesis

If p value $\leq$ alpha: reject the null hypothesis

3 Statistical Inference

The sample variance estimates the population variance, sample variance is a distribution,which centered at the population variance. As we collect more data, the distribution of the sample variance gets more concentrated about population variance.

The sample mean estimates the population mean. Sample mean is a distribution centered at the population mean. It gets more concentrated around the population mean with larger sample size.

The variance of the sample mean is the population variance divided by $n^2.Var(\bar{(}x)) = \frac{\sigma^2}{n}$. The square root is the standard error. The logical estimate of the variance of the sample mean is S^2/n, the logical estimate of hte sample standard error is $S/\sqrt{(}n)$

4 Distributions

4.1 Bernoulli Distribution

1. mean of a Bernoulli random variable is p
2. the variance is p(1-p)
3. the probability mass function$P(X = x) = p^x(1-p)^{1-x}$

4.2 Bi-nominal Distribution

Binominal variable is obtained as the sum of a bunch of iid Bernoulli random variable. In specific, let $X_1, ...X_n$ be iid Bernoulli(p); then $X = \sum_{i=1}^{n} X_i$ is a binomial random variable.

1. P(X=x) = $C_n^x p^x (1-p)^(n-x)$

4.3 Noraml Distribution $X \sim N(\mu, \sigma^2)$

1. Probability density function $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
2. $E(X) = \mu$,$Var(X) = \sigma^2$
3. the area under the f(X) curve between $\mu-$ and $\mu + \sigma$ is 68%, the area under the f(X) curve between $\mu - 2$ and $\mu + 2\sigma$ is 95%
4. $-1.28, -1.645, -1.96, -2.33$ are the $10^{th}, 5^{th}, 2.5^{th}, and 1^{st}$ percentiles of the standard normal distribution respectively.
5. if $X \sim N(\mu, \sigma^2)$, then $Z = \frac{X-\mu}{\sigma} \sim N(0, 1)$
6. Probability area are the same for non standard normal and standard normal, in other words, all normal distributions are identical, the only thing changes is the unit along the axis.

4.4 The Possion distribution: used to model counts

1. $P(X = x, \lambda) = \frac{\lambda^x e - \lambda}{x!}$
2. $E(X) = \lambda$, $Var(X) = \lambda$
3. when n is large and p is small, the Poisson distribution is an accurate approximation to the binomial distribution, $\lambda = np$

5 law of large numbers

It describes the result of performing the same experiment(statistical experiment, not physics experiment) a large number of times. The average of the results obtained from a large number of trials should be close to the expected value, and will tend to become closer as more trials are performed.

It states that the averages of iid samples. converge to the population means that they are estimating

6 central limit theorem

In some situations, when independent random variables are added, their properly normalized sum tends toward a normal distribution even if the original variable themselves are not normally distributed.

suppose that a sample is obtained containing a large number of observations, each observation being randomly generated in a way that does not depend on the values of the other observations, and that the arithmetic average of the observed values is computed. If this procedure is performed many times, the central limit theorem says that the computed values of the average will be distributed according to a normal distribution.

The sample average is approximately $N(\mu, \sigma^2/n)$,μ is the population mean.

7 confidence interval

There's no way for us to know the true parameter of a population, so, we can take an estimate by taking a sample size. We take n samples, and calculate the statistic based on the sample, besides, we can construct an confidence interval about that statistics, the confidence interval will look like: $thevalueofstatistc \pm z^* \frac{\sigma}{\sqrt{n}} error$ but σ is not known, so we use s to estimate σ, then when n is ¡ 30, we got the t distribution, not z/normal distribution If we able to repeatedly get sample of size n from this population, construct a confidence interval in each case, about 95% of the intervals we obtained will contain μ

8 standard error

Sample distribution of an estimator

Variance:

$$\sigma^2 = E[(x-\mu)^2] = E[X^2] - E[x]^2 \tag{1}$$

square root of the variance is called the **standard deviation**.

sample Variance:

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

The variance of the sample mean:

$$Var(\bar{X}) = \frac{\sigma^2}{n}$$

the square root of the variance of the sample mean is called the standard error/standard deviation. Standard error is the standard deviation of its sampling distribution.The sample distribution of a population mean is generated by repeated sampling and recording of the means obtained. This forms a distribution of different means, and this distribution has its own mean and variance.