ME 102: Lecture 17

S. Gopalakrishnan

October 13th, 2016

Confidence Interval for the mean of a Bernoulli Random Variable

Consider a population of items, each of which independently meets certain standards with some unknown probability p. If we let X denote the number of the n items that meet the standards, then X is a binomial random variable with parameters n and p. Thus, when n is large, it follows by the normal approximation to the binomial that X is approximately normally distributed with mean np and variance $np(1-p)$. Hence,

$$\frac{X - np}{\sqrt{np(1-p)}} \sim \mathcal{N}(0,1)$$

where $\sim$ means "is approximately distributed as". Therefore,for any $\alpha \in (0,1)$,

$$P\left\{-z_{\alpha/2} < \frac{X - np}{\sqrt{np(1-p)}} < z_{\alpha/2}\right\} \approx 1 - \alpha$$

Confidence Interval for the mean of a Bernoulli Random Variable

Consider a population of items, each of which independently meets certain standards with some unknown probability p. If we let X denote the number of the n items that meet the standards, then X is a binomial random variable with parameters n and p. Thus, when n is large, it follows by the normal approximation to the binomial that X is approximately normally distributed with mean np and variance $np(1-p)$. Hence,

$$\frac{X - np}{\sqrt{np(1-p)}} \sim \mathcal{N}(0, 1)$$

where $\sim$ means "is approximately distributed as". Therefore,for any $\alpha \in (0, 1)$,

$$P\left\{-z_{\alpha/2} < \frac{X - np}{\sqrt{np(1-p)}} < z_{\alpha/2}\right\} \approx 1 - \alpha$$

and so if X is observed to equal x, then an approximate $100(1-\alpha)$ percent confidence region for p is

$$\left\{p : -z_{\alpha/2} < \frac{x - np}{\sqrt{np(1-p)}} < z_{\alpha/2}\right\}$$

To obtain a confidence interval for p, let $\hat{p} = X/n$ be the fraction of the items that meet the standards. $\hat{p}$ is the maximum likelihood estimator of p, and so should be approximately equal to p. As a result $n\hat{p}(1-\hat{p})$ will be approximately equal to $np(1-p)$

$$\frac{X - np}{\sqrt{n\hat{p}(1-\hat{p})}} \sim \mathcal{N}(0,1)$$

Hence, for any $\alpha \in (0,1)$ we have that

$$P\left\{-z_{\alpha/2} < \frac{X - np}{\sqrt{n\hat{p}(1-\hat{p})}} < z_{\alpha/2}\right\} \approx 1 - \alpha$$

or, equivalently,

$$P\left\{-z_{\alpha/2}\sqrt{n\hat{p}(1-\hat{p})} < np - X < z_{\alpha/2}\sqrt{n\hat{p}(1-\hat{p})}\right\} \approx 1 - \alpha$$

To obtain a confidence interval for p, let $\hat{p} = X/n$ be the fraction of the items that meet the standards. $\hat{p}$ is the maximum likelihood estimator of p, and so should be approximately equal to p. As a result $n\hat{p}(1-\hat{p})$ will be approximately equal to $np(1-p)$

$$\frac{X - np}{\sqrt{n\hat{p}(1-\hat{p})}} \sim \mathcal{N}(0,1)$$

Hence, for any $\alpha \in (0,1)$ we have that

$$P\left\{-z_{\alpha/2} < \frac{X - np}{\sqrt{n\hat{p}(1-\hat{p})}} < z_{\alpha/2}\right\} \approx 1 - \alpha$$

or, equivalently,

$$P\left\{-z_{\alpha/2}\sqrt{n\hat{p}(1-\hat{p})} < np - X < z_{\alpha/2}\sqrt{n\hat{p}(1-\hat{p})}\right\} \approx 1 - \alpha$$

Since $\hat{p} = X/n$, the preceding can be written as

$$P\left\{\hat{p} - z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n} < p < \hat{p} + z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}\right\} \approx 1 - \alpha$$

which yields an approximate $100(1-\alpha)$ percent confidence interval for p.

The Times of India reported that a recent poll indicated that 52 percent of the population was in favor of the job performance of Prime minister Modi, with a margin of error of ± 4 percent. What does this mean? Can we infer how many people were questioned?

The Times of India reported that a recent poll indicated that 52 percent of the population was in favor of the job performance of Prime minister Modi, with a margin of error of ± 4 percent. What does this mean? Can we infer how many people were questioned?

Solution: Assuming the media used a 95 percent confidence interval. The percentage of the population that is in favor of Prime Minister's job performance, is given by

$$\hat{p} \pm 1.96\sqrt{\hat{p}(1-\hat{p}/n} = 0.52 \pm 1.96\sqrt{0.52(0.48)/n}$$

where n is the size of the sample. Since the margin of error is ± 4 percent, it follows that

$$1.96\sqrt{0.52(0.48)/n} = 0.04$$

$$n = 599.29$$

That is, approximately 599 people were sampled, and 52 percent of them reported favorably.

We often want to specify an approximate $100(1-\alpha)$ percent confidence interval for p that is no greater than some given length, say b. The problem is to determine the appropriate sample size n to obtain such an interval.
The length of the approximate $100(1-\alpha)$ percent confidence interval for p from a sample of size n is

$$2z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}$$

which is approximately equal to $2z_{\alpha/2}\sqrt{p(1-p)/n}$. Since p is not known its not possible to directly compute n. We first take a preliminary sample to obtain a rough estimate of p, and then use this estimate to determine n.

$$2z_{\alpha/2}\sqrt{p^*(1-p^*)/n} = b$$

Therefore,

$$n = \frac{(2z_{\alpha/2})^2 p^*(1-p^*)}{b^2}$$

That is, if k items were initially sampled to obtain the preliminary estimate of p, then an additional $n-k$ (or 0 if $n \leq k$) items should be sampled.

A certain manufacturer produces computer chips; each chip is independently acceptable with some unknown probability p. To obtain an approximate 99 percent confidence interval for p, whose length is approximately .05, an initial sample of 30 chips has been taken. If 26 of these chips are of acceptable quality, then the preliminary estimate of p is $26/30$. Using this value, a 99 percent confidence interval of length approximately .05 would require an approximate sample of size

$$n = \frac{4(z_{0.005})^2}{(0.05)^2} \frac{26}{30} \left(1 - \frac{26}{30}\right) \approx 1231$$

A certain manufacturer produces computer chips; each chip is independently acceptable with some unknown probability p. To obtain an approximate 99 percent confidence interval for p, whose length is approximately .05, an initial sample of 30 chips has been taken. If 26 of these chips are of acceptable quality, then the preliminary estimate of p is $26/30$. Using this value, a 99 percent confidence interval of length approximately .05 would require an approximate sample of size

$$n = \frac{4(z_{0.005})^2}{(0.05)^2} \frac{26}{30} \left(1 - \frac{26}{30}\right) \approx 1231$$

Hence, we should now sample an additional 1,201 chips and if, for instance, 1,040 of them are acceptable, then the final 99 percent confidence interval for p is

$$\left(\frac{1066}{1231} - \sqrt{1066\left(1 - \frac{1066}{1231}\right)} \frac{z_{0.005}}{1231}, \frac{1066}{1231} + \sqrt{1066\left(1 - \frac{1066}{1231}\right)} \frac{z_{0.005}}{1231}\right)$$

or

$$p \in (0.84091, 0.89101)$$

As shown, a $100(1-\alpha)$ percent confidence interval for p will be of approximate length b when the sample size is

$$n = \frac{(2z_{\alpha/2})^2}{b^2} p(1-p)$$

$g(p) = p(1-p)$ has maximum value of $\frac{1}{4}$ in the interval $(0,1)$ when $p = \frac{1}{2}$. Thus upper bound on n is

$$n \leq \frac{(z_{\alpha/2})^2}{b^2}$$

and so by choosing a sample whose size is at least as large as $\frac{(z_{\alpha/2})^2}{b^2}$, one can be assured of obtaining a confidence interval of length no greater than b without need of any additional sampling.

Confidence interval of the mean of the Exponential Distribution

If $X_1, X_2, \ldots, X_n$ are independent exponential random variables each having mean θ, then the maximum likelihood estimator of θ is the sample mean $\sum_{i=1}^{n} \frac{X_i}{n}$.
Now, $\sum_{i=1}^{n} X_i$ has a gamma adistribution with parameter $n, 1/\theta$. This implies that

$$\frac{2}{\theta} \sum_{i=1}^{n} X_i \sim \chi^2_{2n}$$

Hence, for any $\alpha \in (0,1)$

$$P\left\{\chi^2_{1-\alpha/2,2n} < \frac{2}{\theta} \sum_{i=1}^{n} X_i < \chi^2_{\alpha/2,2n}\right\} = 1-\alpha$$

or

$$P\left\{\frac{2\sum_{i=1}^{n} X_i}{\chi^2_{\alpha/2,2n}} < \theta < \frac{2\sum_{i=1}^{n} X_i}{\chi^2_{1-\alpha/2,2n}}\right\} = 1-\alpha$$

Hence , a $100(1-\alpha)$ percent confidence interval for θ is

$$\theta \in \left(\frac{2\sum_{i=1}^{n} X_i}{\chi^2_{\alpha/2,2n}}, \frac{2\sum_{i=1}^{n} X_i}{\chi^2_{1-\alpha/2,2n}}\right)$$

The successive items produced by a certain manufacturer are assumed to have useful lives that (in hours) are independent with a common density function

$$f(x) = \frac{1}{\theta} e^{-\frac{x}{\theta}}, \qquad 0 < x < \infty$$

If the sum of the lives of the first 10 items is equal to 1,740, what is a 95 percent confidence interval for the population mean θ?

The successive items produced by a certain manufacturer are assumed to have useful lives that (in hours) are independent with a common density function

$$f(x) = \frac{1}{\theta} e^{-\frac{x}{\theta}}, \qquad 0 < x < \infty$$

If the sum of the lives of the first 10 items is equal to 1,740, what is a 95 percent confidence interval for the population mean θ?

Solution:

$$\chi^2_{0.025,20} = 34.169, \qquad \chi^2_{0.975,20} = 9.661$$

so with 95 percent confidence,

$$\theta \in \left(\frac{3480}{34.169}, \frac{3480}{9.661} \right)$$

or

$$\theta \in (101.847, 360.211)$$

Hypothesis Testing

- A statistical hypothesis is usually a statement about a set of parameters of a population distribution.
- It is called a hypothesis because it is not known whether or not it is true.

Hypothesis Testing

- A statistical hypothesis is usually a statement about a set of parameters of a population distribution.
- It is called a hypothesis because it is not known whether or not it is true.
- A primary problem is to develop a procedure for determining whether or not the values of a random sample from this population are consistent with the hypothesis.

Hypothesis Testing

- A statistical hypothesis is usually a statement about a set of parameters of a population distribution.
- It is called a hypothesis because it is not known whether or not it is true.
- A primary problem is to develop a procedure for determining whether or not the values of a random sample from this population are consistent with the hypothesis.
- For instance, consider a particular normally distributed population having an unknown mean value θ and known variance 1. The statement "θ is less than 1" is a statistical hypothesis that we could try to test by observing a random sample from this population.

Hypothesis Testing

- A statistical hypothesis is usually a statement about a set of parameters of a population distribution.
- It is called a hypothesis because it is not known whether or not it is true.
- A primary problem is to develop a procedure for determining whether or not the values of a random sample from this population are consistent with the hypothesis.
- For instance, consider a particular normally distributed population having an unknown mean value θ and known variance 1. The statement "θ is less than 1" is a statistical hypothesis that we could try to test by observing a random sample from this population.
- If the random sample is deemed to be consistent with the hypothesis under consideration, we say that the hypothesis has been "accepted"; otherwise we say that it has been "rejected".

Significance levels

- Consider a population having distribution F_θ , where θ is unknown, and suppose we want to test a specific hypothesis about θ.
- We shall denote this hypothesis by H_0 and call it the null hypothesis.

Significance levels

- Consider a population having distribution F_θ , where θ is unknown, and suppose we want to test a specific hypothesis about θ.
- We shall denote this hypothesis by H_0 and call it the null hypothesis.
- For example, if F_θ is a normal distribution function with mean θ and variance equal to 1, then two possible null hypotheses about θ are
  (a) $H_0 : \theta = 1$
  (b) $H_0 : \theta \leq 1$
- Thus the first of these hypotheses states that the population is normal with mean 1 and variance 1, whereas the second states that it is normal with variance 1 and a mean less than or equal to 1.
- A hypothesis that, when true, completely specifies the population distribution is called a **simple** hypothesis; one that does not is called a **composite** hypothesis.

A test for H_0 can be specified by defining a region C in n–dimensional space with the proviso that the hypothesis is to be rejected if the random sample $X_1, \ldots, X_n$ turns out to lie in C and accepted otherwise. The region C is called the critical region. In other words, the statistical test determined by the critical region C is the one that

$$\text{accepts } H_0 \quad \text{if } (X_1, \ldots, X_n) \notin C$$

and

$$\text{rejects } H_0 \quad \text{if } (X_1, \ldots, X_n) \in C$$

A test for H_0 can be specified by defining a region C in n–dimensional space with the proviso that the hypothesis is to be rejected if the random sample $X_1, \ldots, X_n$ turns out to lie in C and accepted otherwise. The region C is called the critical region. In other words, the statistical test determined by the critical region C is the one that

$$\text{accepts } H_0 \quad \text{if } (X_1, \ldots, X_n) \notin C$$

and

$$\text{rejects } H_0 \quad \text{if } (X_1, \ldots, X_n) \in C$$

For instance, a common test of the hypothesis that θ, the mean of a normal population with variance 1, is equal to 1 has a critical region given by

$$C = \left\{ (X_1, \ldots, X_n) : \left| \frac{\sum_{i=1}^{n} X_i}{n} - 1 \right| > \frac{1.96}{\sqrt{n}} \right\}$$

Thus, this test calls for rejection of the null hypothesis that $\theta = 1$ when the sample average differs from 1 by more than 1.96 divided by the square root of the sample size.

Errors

- When developing a procedure for testing a given null hypothesis H_0 that, in any test, two different types of errors can result.
- The first of these, called a type I error, is said to result if the test incorrectly calls for rejecting H_0 when it is indeed correct.
- The second, called a type II error, results if the test calls for accepting H_0 when it is false.

Errors

- When developing a procedure for testing a given null hypothesis H_0 that, in any test, two different types of errors can result.
- The first of these, called a type I error, is said to result if the test incorrectly calls for rejecting H_0 when it is indeed correct.
- The second, called a type II error, results if the test calls for accepting H_0 when it is false.
- The objective of a statistical test of H_0 is not to explicitly determine whether or not H_0 is true but rather to determine if its validity is consistent with the resultant data.
- Hence, with this objective it seems reasonable that H_0 should only be rejected if the resultant data are very unlikely when H_0 is true.

Errors

- When developing a procedure for testing a given null hypothesis H_0 that, in any test, two different types of errors can result.
- The first of these, called a type I error, is said to result if the test incorrectly calls for rejecting H_0 when it is indeed correct.
- The second, called a type II error, results if the test calls for accepting H_0 when it is false.
- The objective of a statistical test of H_0 is not to explicitly determine whether or not H_0 is true but rather to determine if its validity is consistent with the resultant data.
- Hence, with this objective it seems reasonable that H_0 should only be rejected if the resultant data are very unlikely when H_0 is true.
- The classical way of accomplishing this is to specify a value α and then require the test to have the property that whenever H_0 is true its probability of being rejected is never greater than α.
- The value α, called the level of significance of the test, is usually set in advance.

TESTS CONCERNING THE MEAN OF A NORMAL POPULATION

Suppose that $X_1, \ldots, X_n$ is a sample of size n from a normal distribution having an unknown mean μ and a known variance σ^2 and suppose we are interested in testing the null hypothesis

$$H_0 : \mu = \mu_0$$

against the alternative hypothesis

$$H_1 : \mu \neq \mu_o$$

where μ_0 is some specified constant.

TESTS CONCERNING THE MEAN OF A NORMAL POPULATION

Suppose that $X_1, \ldots, X_n$ is a sample of size n from a normal distribution having an unknown mean μ and a known variance σ^2 and suppose we are interested in testing the null hypothesis

$$H_0 : \mu = \mu_0$$

against the alternative hypothesis

$$H_1 : \mu \neq \mu_o$$

where μ_0 is some specified constant. Since $\bar{X} = \sum_{i=1}^{n} X_i/n$ is a point estimator of μ, it seems reasonable to accept H_0 if $\bar{X}$ is not too far from μ_0. The critical region for the test would be,

$$C = \{X_1, \ldots, X_n : |\bar{X} - \mu_0| > c\}$$

for some suitably chosen value c. If we desire that the test has significance level α, c must be such that

$$P_{\mu_0}\{|\bar{X} - \mu_0| > c\} = \alpha$$

where we write P_{μ_0} to mean that the preceding probability is to be computed under the assumption that $\mu = \mu_0$.

However, when $\mu = \mu_0$, $\bar{X}$ will be normally distributed with mean μ_0 and variance σ^2/n and so Z, defined by

$$Z \equiv \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

will have a standard normal distribution.

However, when $\mu = \mu_0$, $\bar{X}$ will be normally distributed with mean μ_0 and variance σ^2/n and so Z, defined by

$$Z \equiv \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

will have a standard normal distribution. hence,

$$P\left\{|Z| > \frac{c\sqrt{n}}{\sigma}\right\} = \alpha$$

or

$$2P\left\{Z > \frac{c\sqrt{n}}{\sigma}\right\} = \alpha$$

However, when $\mu = \mu_0$, $\bar{X}$ will be normally distributed with mean μ_0 and variance σ^2/n and so Z, defined by

$$Z \equiv \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

will have a standard normal distribution. hence,

$$P\left\{|Z| > \frac{c\sqrt{n}}{\sigma}\right\} = \alpha$$

or

$$2P\left\{Z > \frac{c\sqrt{n}}{\sigma}\right\} = \alpha$$

We know that,

$$P\{Z > z_{\alpha/2}\} = \alpha/2$$

and so

$$\frac{c\sqrt{n}}{\sigma} = z_{\alpha/2} \quad c = \frac{z_{\alpha/2}\sigma}{\sqrt{n}}$$

Thus, for the significance level α test

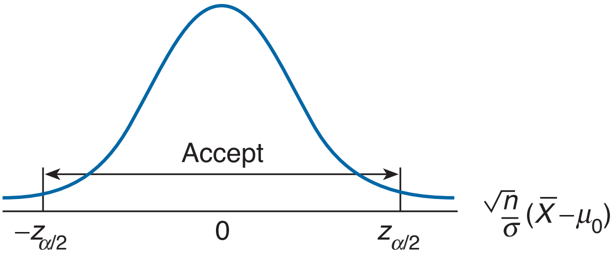
$$\text{reject } H_0 \quad \text{if } \frac{\sqrt{n}}{\sigma}|\bar{X} - \mu_0| > z_{\alpha/2}$$

$$\text{accept } H_0 \quad \text{if } \frac{\sqrt{n}}{\sigma}|\bar{X} - \mu_0| \leq z_{\alpha/2}$$

Thus, for the significance level α test

$$\text{reject } H_0 \quad \text{if } \frac{\sqrt{n}}{\sigma}|\bar{X} - \mu_0| > z_{\alpha/2}$$

$$\text{accept } H_0 \quad \text{if } \frac{\sqrt{n}}{\sigma}|\bar{X} - \mu_0| \leq z_{\alpha/2}$$



It is known that if a signal of value μ is sent from location A, then the value received at location B is normally distributed with mean μ and standard deviation 2. That is, the random noise added to the signal is an $N(0, 4)$ random variable. There is reason for the people at location B to suspect that the signal value $\mu = 8$ will be sent today. Test this hypothesis if the same signal value is independently sent five times and the average value received at location B is $\bar{X} = 9.5$.

It is known that if a signal of value μ is sent from location A, then the value received at location B is normally distributed with mean μ and standard deviation 2. That is, the random noise added to the signal is an $N(0, 4)$ random variable. There is reason for the people at location B to suspect that the signal value $\mu = 8$ will be sent today. Test this hypothesis if the same signal value is independently sent five times and the average value received at location B is $\bar{X} = 9.5$.

Suppose we are testing at the 5 percent level of significance. To begin, we compute the test statistic

$$\frac{\sqrt{n}}{\sigma}|\bar{X} - \mu| = \frac{\sqrt{5}}{2}(1.5) = 1.68$$

Since this value is less than $z_{.025} = 1.96$, the hypothesis is accepted. In other words, the data are not inconsistent with the null hypothesis in the sense that a sample average as far from the value 8 as observed would be expected, when the true mean is 8, over 5 percent of the time.