Stat 400 : Discussion section BD3 and BD4 Handout 9

Subhadeep Paul

April 16, 2013

1 CI for mean μ, variance σ and proportion p

- When σ^2 is unknown, we simply estimate it with $s^2 = \frac{1}{n-1}\sum(x_i - \bar{x})^2$. We could use $\frac{1}{n}\sum(x_i - \bar{x})^2$ as well as an estimate which is the usual expression of variance, but this estimate is *biased* (the proof is simple and try it as an exercise). So we use s^2 , which is unbiased estimate of σ^2. This is called *sample variance*. It follows chi squared distribution with $n-1$ degrees of freedom.

Problem 1. (midterm 3, fall 12) Studies have shown that more than 12 million tin-coated steel cans are removed from the municipal waste streams of our cities and recycled each day. Suppose it is desired to estimate the mean number of tin cans recovered from mixed refuse per year in American cities. A random sample of 9 American cities yielded the following summary statistics on number of tin cans (in millions) recovered per city last year: $\bar{y} = 110$, $s = 10.5$. Construct a 95% confidence interval for the true mean number of tin cans removed annually from mixed refuse for recycling in American cities.

- ***Confidence interval for*** σ^2 ***is***

$$(\frac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}}, \frac{(n-1)S^2}{\chi^2_{n-1,(1-\alpha/2)}})$$

- ***Confidence interval for functions of parameters*** If $f(\theta)$ is a function of θ and CI for θ is $c(x)$, then CI for $f(\theta)$ is $f(c(x))$. So CI of a function of parameter is obtained by simply applying the function to the CI of the parameter. e.g σ is square root of σ^2. So, CI of σ will be square root of the CI for σ.

- ***Minimum length*** For 95% CI of mean with known variance consider these two intervals, both of them are 95%, one symmetric about $\bar{X}$, $\pm 1.96\frac{\sigma}{\sqrt{n}}$ and the other one, $(\bar{X} - 1.66\frac{\sigma}{\sqrt{n}}, \bar{X} + 2.81\frac{\sigma}{\sqrt{n}})$. The 2nd one is an bigger interval and hence is less precise. So we always choose the one with minimum length. For normal its the symmetric CI. For chi square pdf (which is not symmetric , find a and b from table) that correspond to minimum length interval.

Problem 2. (final, fall 12) A graduate admissions office is analyzing data about a random sample of 22 applicants' GRE scores. The the sample has a mean of 600 and standard deviation of 100. (a) Find a 95% lower-bound confidence interval for the overall standard deviation, (b) Construct the 90% minimum length confidence interval for the overall standard deviation.

Table 1: CI for mean and proportions

	σ known	σ unknown	proportion $\hat{p}$
Point estimate	$\bar{X}$	$\bar{X}$	$\hat{p} = \frac{x}{n}$
standard error	$\frac{\sigma}{\sqrt{n}}$	$\frac{s}{\sqrt{n}}$	$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
distribution	Normal	t	Normal
total error term ϵ	$z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$	$t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}$	$z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
sample size	$(\frac{\sigma}{\epsilon}z_{\alpha/2})^2$	$(\frac{s}{\epsilon}t_{n-1,\alpha/2})^2$	$(\frac{p*(1-p*)}{\epsilon}z_{\alpha/2})^2$

Table 2: Hypothesis testing

	Accept	*Reject*
H_0	OK	Type 1
H_1	Type 2	OK

Court judgement	Hypothesis testing
Defendent innocent	Null is true
Evidence	data
Defendent declared guilty	Null rejected
Defendent not guilty	Fail to reject null
Reasonable doubt	Level of significance

- ***Proportions*** See the table for formulae. We estimate the unknown population proportion from (observed) sample proportion as $\hat{p} = \frac{x}{n}$. With Binomial assumption, $E(\hat{p}) = p$ and $var(\hat{p}) = \frac{p(1-p)}{n}$. For sample size calculation $p*$ is either taken as the value specified by the null hypothesis or 0.50.

Problem 3. (midterm 3, fall 2012)A polling firm conducts a poll to determine what proportion p of voters in a given population will vote in an upcoming election. A random sample of 250 was taken from the population, and the proportion answering yes was 0.66. Find a 98% confidence interval for proportion of voters endorsing the candidate in the population.

2 Hypothesis testing

2.1 Null and alternative, link with CI (courtsey: Prof John Marden, stat 511 notes)

- Estimation addresses the question, "What is ϑ?" Hypothesis testing addresses questions like, "Is $\vartheta = 0$?" Confidence intervals do both. It will give a range of plausible values, and if you wonder whether $\vartheta = 0$ is plausible. you just check whether 0 is in the interval.

- Null hypothesis is something that you start your test with. You don't want to reject it unless there is substantial evidence from data against the null (and hence in favor of alternative). If the evidence is overwhelming and beyond reasonable doubt, you will reject the null or else you will fail to reject (accept) the null. So we don't want to reject our null by error. Type 1 error is worse than type 2 error. Now the problem is you can't reduce both type 1 and type 2 error at the same time. We must first control for Type 1 error and then try to minimize type 2 error.

- P(Type 1 error)=α, the level of significance, 1-P(Type 2 error)=power. The test statistic is a function of data $\phi(X)$. The values of the test statistic for which we reject the null is called rejection region.

Problem 4. Just prior to an important election, in a random sample of 749 voters, 397 preferred Candidate Y over Candidate Z. Is there enough evidence to claim that over half of all the voters prefer Candidate Y over Candidate Z? (a) What hypothesis are we interested in testing? Write Null and alternative (b) What is the value of the appropriate test statistic? (c)What is the rejection region? (d) Perform an appropriate test at a 10% level of significance. (e) Find the p-value of the appropriate test.

2.2 p-value: The most used statistics phrase outside statistics. Interpretation

Probability of obtaining as extreme a test statistic as the one observed by sheer chance. Again p-value doesn't directly say anything about the result of the test but says about the significance of the testing procedure. For our purpose, we need to report the probability value corresponding to the test statistic from table