Confidence Intervals - Class 8

January 30, 2014 Debdeep Pati

Estimation

1 Interval Estimation

1. Quantify the uncertainty
2. The 10 birthweigths $97, 125, 62, 120, 132, 135, 118, 137, 126, 118$ have a mean of 116.9 oz. How certain are we that the true mean is 116.9 oz? $116.9oz \pm 1$ oz and $116.9oz \pm 1$ lb are certainly different.
3. The sample mean $\bar{X} \sim N(\mu, \sigma^2/n)$. If μ and σ^2 are known then if you keep on generating samples, 95% of all such samples will fall in the interval

$$(\mu - 1.96\sigma/\sqrt{n}, \mu + 1.96\sigma/\sqrt{n})$$

4. We can also express the mean in standardized form by

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

5. 95% of Z value from repeated samples of size n will fall between -1.96 and +1.96.
6. However, the assumption that σ is known is quite artificial. Since σ is unknown, we can estimate σ by the sample standard deviation s and construct confidence intervals using

$$\frac{\bar{X} - \mu}{s/\sqrt{n}}$$

7. This quantity is no longer normally distributed. The distribution is called Students t distribution, or t distribution if X_i's are normally distributed. t distribution is not a unique distribution. It is a family of distributions indexed by a parameter, the degrees of freedom (df).
8. If $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ and are independent, then $\frac{\bar{X}-\mu}{s/\sqrt{n}}$ is distributed as a t distribution with $n - 1$ degrees of freedom.

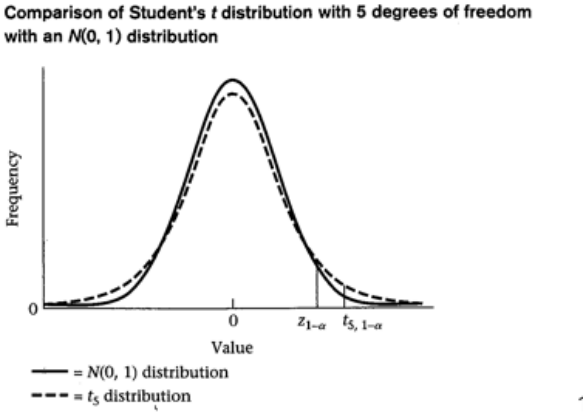
9. The $100 \times u$ th percentile of a t distribution is d degrees of freedom is denoted by $t_{d,u}$, that is,

$$P(t_d < t_{d,u}) = u$$

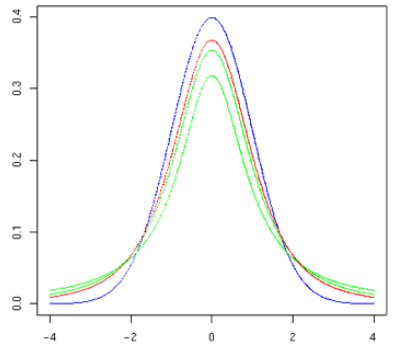
10. $t_{20,0.95}$ stands for the95 th percentile or the upper 5th percentile of a t distribution with 20 degrees of freedom.

1.1 Comparison of t and Normal distribution

1. Compare t with d degrees of freedom to N(0, 1). For any $\alpha > 0.5$,$t_{d,1-\alpha}$ is always larger than the corresponding percentile for an $N(0,1)(z_{1-\alpha})$. When d becomes large t converge to $N(0,1)$.



- T distribution with degree of freedom 1, 2, 3 and Normal (blue).



2. Find the upper 5th percentile of a t distribution with 23 df. $t_{23,.95}$ is given in row 23 and column 0.95 of Table 5 and is 1.714. Probabilities associated with t distribution can also be calculated using statistical programs. In $R : qt(.95, 23)$

3. If σ is unknown we want $100(1-\alpha)\%$ of the t statistics should fall between the lower and upper quantile of a t_{n-1} distribution.

$$P(t_{n-1,\alpha/2} < t < t_{n-1,1-\alpha/2}) = 1 - \alpha$$

The above equation leads to

$$P(\bar{X} - t_{n-1,\alpha/2}S/\sqrt{n} < \mu < \bar{X} + t_{n-1,\alpha/2}S/\sqrt{n}) = 1 - \alpha$$

The interval $[\bar{X} - t_{n-1,\alpha/2}S/\sqrt{n}, \bar{X} + t_{n-1,\alpha/2}S/\sqrt{n}]$ is referred to as a $100(1-\alpha)\%$ confidence interval for μ.

4. Compute a 95% CI for the mean birthweight based on the sample of size 10 in the previous example.

We have $n = 10, \bar{x} = 116.90, s = 21.70$.
Because we want 95% CI, $\alpha = .05$.
The 95% CI is
$[116.90 - t_{9,.975}(21.70)/\sqrt{10}, 116.90 + t_{9,.975}(21.70)/\sqrt{10}]$
From Table 5, $t_{9,.975} = 2.262$. (qt(.975,9) in R)
95% CI is
$[116.90 - 2.262(21.70)/\sqrt{10}, 116.90 + 2.262(21.70)/\sqrt{10}]$
$= (116.90 - 15.5, 116.90 + 15.5)$
$= (101.4, 132.4)$

5. Using confidence interval in decision making: Suppose we know the mean cholesterol level in children ages 2-14 is 175 mg/dL. We wish to see if there is a familial aggregation of cholesterol levels. Identify a group of fathers with cholesterol levels $\geq 250 mg/dL$ and measure the cholesterol levels of their 2-14-year-old offspring. Suppose we find the mean cholesterol level in a group of 100 such children is 207.3 mg/dL with standard deviation = 30 mg/dL. Is this value far enough from 175 mg/dL for us to believe that the underlying mean cholesterol level in the population of all children is different from 175 mg/dL?
Construct a 95% CI for μ on the basis of our sample data. Decision rule: if the interval contains 175 mg/dL, then we cannot say the underlying mean for this group is any different from the mean for all children (175). If the CI does not contain 175, then we would conclude the true underlying mean for this group is different from 175. If the lower bound of the CI is above 175, then there is a demonstrated familial aggregation of cholesterol levels. The CI in this case is given by

$$207.3 \pm t_{99,0.975}(30)/\sqrt{100} = 207.3 \pm 6.0 = (201.3, 213.3)$$

2 Case study

1. Assess whether there is a relationship between bone-mineral density (BMD) and cigarette smoking. 41 twin pairs are selected and each pair has different smoking

histories. Matched-pair study - The exposed (heavier-smoking twin) and control (lighter-smoking twin) are matched on other characteristics related to the outcome (BMD). The matching is based on having similar genes so that the effect of genes on outcome can be safely ignored.

2. The difference in bone-mineral density (BMD) was studied as function of the difference in tobacco use. Difference: BMD of heavier-smoking twin - BMD of lighter-smoking twin. Tobacco consumption was expressed in terms of pack-year. One pack-year is defined as 1 pack of cigarettes per day consumed for a year. BMD was assessed separately at three sites - The lumbar spine, the femoral neck, and the femoral shaft.

3. To assess whether there is a relationship between BMD and cigarette smoking. Calculate the difference in BMD between heavier-smoking twin and the lighter-smoking twin for each twin pair. Calculate the average of these differences, which is $-0.036 \pm 0.014 g/cm^2$.

4. The 95% CI for the true mean difference in BMD is

$$-0.036 \pm t_{40,0.975}(s/\sqrt{41}) = -0.036 \pm 2.021(0.014) = (-0.064, -0.008)$$

5. Upper bound is less than 0. The true mean difference is less than 0. The true mean BMD for the heavier-smoking twins is lower than for the lighter-smoking twins.

Estimation of the Variance - point estimate

1. Definition of sample variance

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

2. Let $X_1, \ldots, X_n$ be a random sample from some population with mean μ and variance σ^2.

3. The sample variance S^2 is an unbiased estimator of σ^2 over all possible random samples of size n that could have been drawn from this population; $E(S^2) = \sigma^2$.

Suppose we have the data in the following table, consisting of systolic blood pressure (SBP) measurements obtained on 10 people and read by two observers. We use the difference d_i between the first and second observers to assess interobserver variability. In particular, if we assume the underlying distribution of these differences is normal with mean μ and

Table 6.6 SBP measurements (mm Hg) from an Arteriosonde machine obtained from 10 people and read by two observers

	Observer		
Person (i)	1	2	Difference (d)
1	194	200	−6
2	126	123	+3
3	130	128	+2
4	98	101	−3
5	136	135	+1
6	145	145	0
7	110	111	−1
8	108	107	+1
9	102	99	+3
10	126	128	−2

We have seen previously that an unbiased estimator of the variance σ^2 is given by the sample variance S^2. In this case,

$$\text{Mean difference} = (-6+3+\cdots-2)/10 = -0.2 = \bar{d}$$

$$\text{Sample variance} = s^2 = \sum_{i=1}^{n}(d_i-\bar{d})^2/9$$

$$= \left[(-6+0.2)^2+\cdots+(-2+0.2)^2\right]/9 = 8.178$$

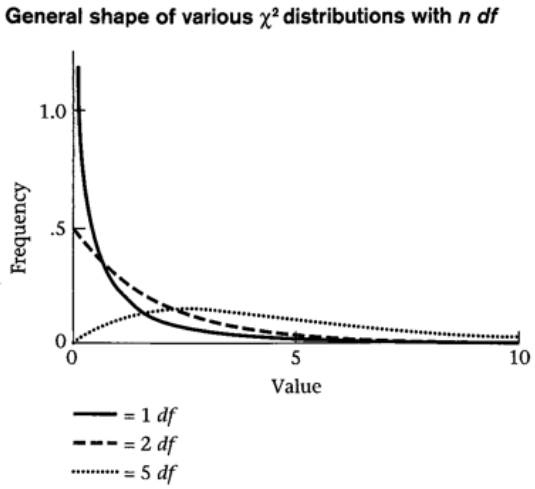
how can an interval estimate for σ^2 be obtained?

variance σ^2, then it is of primary interest to estimate σ^2. The higher σ^2 is, the higher the interobserver variability.

We could use the CI for σ^2 to make decisions concerning the variability of the Arteriosonde machine if we had a good estimate of the interobserver variability of blood-pressure readings from a standard cuff. For example, suppose we know from previous work that if two people are listening to blood-pressure recordings from a standard cuff, then the interobserver variability as measured by the variance of the set of differences between the readings of two observers is 35. This value is outside the range of the 95% CI for σ (3.87, 27.26), and we thus conclude that the inter-observer variability is reduced by using an Arteriosonde machine. Alternatively, if this prior variance were 15, then we could not say that the variances obtained from using the two methods are different.

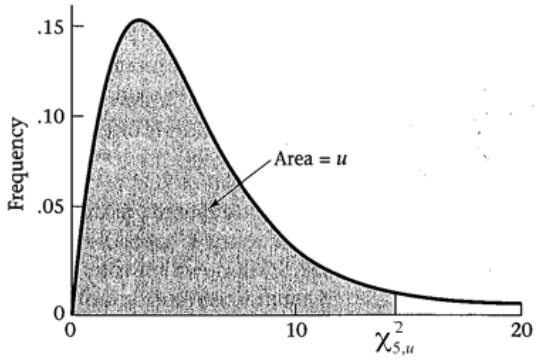
Interval estimation for σ^2

1. If $G = \sum_{i=1}^{n} X_i^2$, where $X_1, X_2, \ldots, X_n \sim N(0,1)$, then G is said to follow a chi-square distribution with n degrees of freedom (df). This is often denoted by χ_n^2.
2. $\frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2$ if $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$.



3. The uth percentile of a χ_n^2 distribution is denoted by $\chi_{n,u}^2$, where $P(\chi_n^2 < \chi_{n,u}^2) = u$.

Figure 6.9 Graphic display of the percentiles of a χ^2_5 distribution



4. Assuming $P\left(\frac{\sigma^2\chi^2_{n-1,\alpha/2}}{n-1} < S^2 < \frac{\sigma^2\chi^2_{n-1,1-\alpha/2}}{n-1}\right)$, a $100(1-\alpha)\%$ confidence interval for σ^2 is

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