Estimation

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 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 $?^2$ (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.

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,

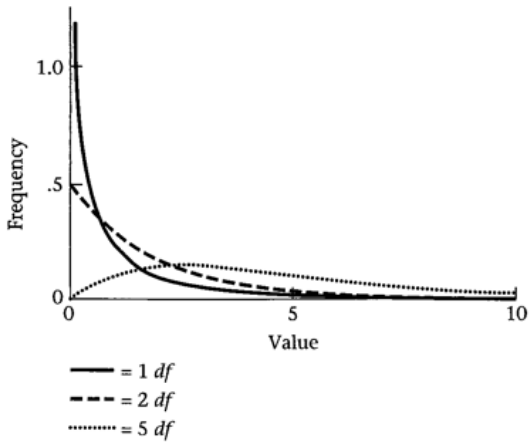
$$\text{Mean difference} = (-6+3+\cdots-2)/10 = -0.2 = \bar{d}$$

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

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

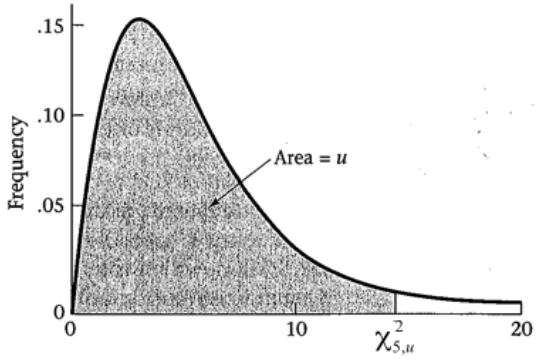
how can an interval estimate for σ^2 be obtained?

General shape of various χ^2 distributions with *n df*



2. $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$ if $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$.
3. The uth percentile of a χ^2_n distribution is denoted by $\chi^2_{n,u}$, where $P(\chi^2_n < \chi^2_{n,u}) = u$.

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

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

Estimation for Binomial distribution

1. Estimating the prevalence of malignant melanoma in 45-54 year old women in the US. A random sample of 5000 women is selected from this age group and 28 are found to have the disease. Let random variable X_i represent the disease status for the ith woman, where $X_i = 1$ if the ith woman has the disease and 0 if she does not. Suppose the prevalence of the disease in this age group is p. How can p be estimated?

2. Let $X \sin \text{Bin}(n, p)$. Then an unbiased of p is X/n. The standard error is $\sqrt{p(1-p)/n}$.

3. Estimate the prevalence of malignant melanoma in the previous example and provide its standard error. The best estimator of the prevalence of the disease is $28/5000 = .0056$. Its estimated standard error is $\sqrt{0.00056(0.9944)/5000} = 0.0011$.

Maximum Likelihood estimation

1. Suppose we have a sample of 100 men of whom 30 have diabetes. If the prevalence of diabetes = p, then what is the likelihood of the sample given p?

2. Suppose we have n independent observations $x_1, \ldots, x_n$ from a normal distribution with mean μ and variance of σ^2. What is the likelihood of the sample?

Estimation for the Binomial Distribution - Interval Estimation

Estimating the prevalence rate of breast cancer among 50- to 54-year-old women whose mothers have had breast cancer. In a random sample of 10,000 such women, 400 are found to have had breast cancer at some point in their lives. The best point estimate of the prevalence rate p is given by the sample proportion p = 400/10,000 = .04. Interval estimate of p? $p \pm z_{1-\alpha/2}\sqrt{p(1-p)/n}$