

## **Topic 9: Sampling Distributions of Estimators**

**Rohini Somanathan**

**Course 003, 2016**

## Sampling distributions of estimators

- Since our estimators are statistics (particular functions of random variables), their distribution can be derived from the joint distribution of  $X_1 \dots X_n$ . It is called the **sampling distribution** because it is based on the joint distribution of the random sample.
- Given a sampling distribution, we can
  - calculate the probability that an estimator will not differ from the parameter  $\theta$  by more than a specified amount
  - obtain **interval estimates** rather than point estimates after we have a sample- an interval estimate is a random interval such that the true parameter lies within this interval with a given probability (say 95%).
  - **choose between to estimators**- we can, for instance, calculate the mean-squared error of the estimator,  $E_{\theta}[(\hat{\theta} - \theta)^2]$  using the distribution of  $\hat{\theta}$ .
- **Sampling distributions of estimators depend on sample size**, and we want to know exactly how the distribution changes as we change this size so that we can make the right trade-offs between cost and accuracy.

## Sampling distributions: sample size and precision

### Examples:

1. What if  $X_i \sim N(\theta, 4)$ , and we want  $E(\bar{X}_n - \theta)^2 \leq .1$ ? This is simply the variance of  $\bar{X}_n$ , and we know  $\bar{X}_n \sim N(\theta, 4/n)$ .

$$\frac{4}{n} \leq .1 \text{ if } n \geq 40$$

2. Consider a random sample of size  $n$  from a **Uniform distribution on  $[0, \theta]$** , and the statistic  $U = \max\{X_1, \dots, X_n\}$ . The CDF of  $U$  is given by:

$$F(u) = \begin{cases} 0 & \text{if } u \leq 0 \\ \left(\frac{u}{\theta}\right)^n & \text{if } 0 < u < \theta \\ 1 & \text{if } u \geq \theta \end{cases}$$

We can now use this to see how large our sample must be if we want a certain level of precision in our estimate for  $\theta$ . Suppose we want the probability that our estimate lies within  $.1\theta$  for any level of  $\theta$  to be bigger than 0.95:

$$\Pr(|U - \theta| \leq .1\theta) = \Pr(\theta - U \leq .1\theta) = \Pr(U \geq .9\theta) = 1 - F(.9\theta) = 1 - 0.9^n$$

We want this to be bigger than 0.95, or  $0.9^n \leq 0.05$ . With the LHS decreasing in  $n$ , we choose  $n \geq \frac{\log(.05)}{\log(.9)} = 28.43$ . Our minimum sample size is therefore 29.

## Joint distribution of sample mean and sample variance

For a **random sample from a normal distribution**, we know that the M.L.E.s are the sample mean and the sample variance  $\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2$ . Also,

- $\bar{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right)$
- $\sum_{i=1}^n \left(\frac{X_i - \mu}{\sigma}\right)^2 \sim \chi_n^2$  (since it is the sum of squares of  $n$  standard normal random variables).
- If we replace the population mean  $\mu$  with the sample mean  $\bar{X}_n$ , the resulting sum of squares, has a  $\chi_{n-1}^2$  distribution.

**Theorem:** If  $X_1, \dots, X_n$  form a random sample from a **normal distribution** with mean  $\mu$  and variance  $\sigma^2$ , then the sample mean  $\bar{X}_n$  and the sample variance  $\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2$  are independent random variables and

$$\bar{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

$$\frac{\sum_{i=1}^n (X_i - \bar{X}_n)^2}{\sigma^2} \sim \chi_{n-1}^2$$

**Note:** This is only for normal samples. Work through the application of this theorem on p. 475 of your textbook, where you are asked to compute the probability that the sample mean and sample standard deviation of a sample drawn from a  $N(\mu, \sigma^2)$  are within  $.2\sigma$  of their population values.

## The t-distribution

Let  $Z \sim N(0,1)$ , let  $Y \sim \chi_v^2$ , and let  $Z$  and  $Y$  be independent random variables. Then

$$X = \frac{Z}{\sqrt{\frac{Y}{v}}} \sim t_v$$

The p.d.f of the **t-distribution** is given by:

$$f(x; v) = \frac{\Gamma(\frac{v+1}{2})}{\Gamma(\frac{v}{2})\sqrt{\pi v}} \left(1 + \frac{x^2}{v}\right)^{-\frac{v+1}{2}}$$

**Features of the t-distribution:**

- One can see from the above density function that the t-density is symmetric with a maximum value at  $x = 0$ .
- The shape of the density is similar to that of the standard normal (bell-shaped) but with fatter tails.

## Relation to random normal samples

**RESULT 1:** Define  $S_n^2 = \sum_{i=1}^n (X_i - \bar{X}_n)^2$  The random variable

$$U = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sqrt{\frac{S_n^2}{n-1}}} \sim t_{n-1}$$

**Proof:** We know that  $\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \sim N(0,1)$  and that  $\frac{S_n^2}{\sigma^2} \sim \chi_{n-1}^2$ . Dividing the first random variable by the square root of the second, divided by its degrees of freedom, the  $\sigma$  in the numerator and denominator cancels to obtain  $U$ .

**Implication:** We cannot make statements about  $|\bar{X}_n - \mu|$  using the normal distribution if  $\sigma^2$  is unknown. This result allows us to use its estimate  $\hat{\sigma}^2 = \sum_{i=1}^n (X_i - \bar{X}_n)^2/n$  since  $\frac{(\bar{X}_n - \mu)}{\hat{\sigma}/\sqrt{n-1}} \sim t_{n-1}$

**RESULT 2** Given  $X, Z, Y, n$  as above. As  $n \rightarrow \infty$   $X \xrightarrow{d} Z \sim N(0,1)$

**To see why:**  $U$  can be written as  $\sqrt{\frac{n-1}{n}} \frac{\sqrt{n}(\bar{X}_n - \mu)}{\hat{\sigma}} \sim t_{n-1}$ . As  $n$  gets large  $\hat{\sigma}$  gets very close to  $\sigma$  and  $\frac{n-1}{n}$  is close to 1.

$F^{-1}(.55) = .129$  for  $t_{10}$ ,  $.127$  for  $t_{20}$  and  $.126$  for the standard normal distribution. The differences between these values increases for higher values of their distribution functions (why?)

## Confidence intervals for the mean

Given  $\sigma^2$ , let us see how we can obtain an **interval estimate** for  $\mu$ , i.e. an interval which is likely to contain  $\mu$  with a pre-specified probability.

- Since  $\frac{(\bar{X}_n - \mu)}{\sigma/\sqrt{n}} \sim \mathbf{N}(0,1)$ ,  $\Pr\left(-2 < \frac{(\bar{X}_n - \mu)}{\sigma/\sqrt{n}} < 2\right) = .955$
- But this event is equivalent to the events  $-\frac{2\sigma}{\sqrt{n}} < \bar{X}_n - \mu < \frac{2\sigma}{\sqrt{n}}$  and  $\bar{X}_n - \frac{2\sigma}{\sqrt{n}} < \mu < \bar{X}_n + \frac{2\sigma}{\sqrt{n}}$
- With known  $\sigma$ , each of the random variables  $\bar{X}_n - \frac{2\sigma}{\sqrt{n}}$  and  $\bar{X}_n + \frac{2\sigma}{\sqrt{n}}$  are statistics. Therefore, we have derived a random interval within which the population parameter lies with probability .955, i.e.

$$\Pr\left(\bar{X}_n - \frac{2\sigma}{\sqrt{n}} < \mu < \bar{X}_n + \frac{2\sigma}{\sqrt{n}}\right) = .955 = \gamma$$

- Notice that there are many intervals for the same  $\gamma$ , this is the shortest one.
- Now, given our sample, our statistics take particular values and the resulting interval either contains or does not contain  $\mu$ . We can therefore no longer talk about the probability that it contains  $\mu$  because the experiment has already been performed.
- We say that  $(\bar{x}_n - \frac{2\sigma}{\sqrt{n}} < \mu < \bar{x}_n + \frac{2\sigma}{\sqrt{n}})$  is a 95.5% **confidence interval** for  $\mu$ . Alternatively, we may say that  $\mu$  lies in the above interval with **confidence**  $\gamma$  or that the above interval is a confidence interval for  $\mu$  with **confidence coefficient**  $\gamma$

## Confidence Intervals for means..examples

- **Example 1:**  $X_1, \dots, X_n$  forms a random sample from a normal distribution with unknown  $\mu$  and  $\sigma^2 = 10$ .  $\bar{x}_n$  is found to be 7.164 with  $n = 40$ . An 80% confidence interval for the mean  $\mu$  is given by  $(7.164 - 1.282\sqrt{\frac{10}{40}}, 7.164 + 1.282\sqrt{\frac{10}{40}})$  or  $(6.523, 7.805)$ . The **confidence coefficient** is .8
- **Example 2:** Let  $\bar{X}$  denote the sample mean of a random sample of size 25 from a distribution with variance 100 and mean  $\mu$ . In this case,  $\frac{\sigma}{\sqrt{n}} = 2$  and, making use of the central limit theorem the following statement is approximately true:

$$\Pr\left(-1.96 < \frac{(\bar{X}_n - \mu)}{2} < 1.96\right) = .95 \text{ or } \Pr\left(\bar{X}_n - 3.92 < \mu < \bar{X}_n + 3.92\right) = .95$$

If the sample mean is given by  $\bar{x}_n = 67.53$ , an approximate 95% confidence interval for the sample mean is given by  $(63.61, 71.45)$ .

- **Example 3:** Suppose we are interested in a confidence interval for the mean of a normal distribution but do not know  $\sigma^2$ . We know that  $\frac{(\bar{X}_n - \mu)}{\hat{\sigma}/\sqrt{n-1}} \sim t_{n-1}$  and can use the t-distribution with  $(n - 1)$  degrees of freedom to construct our interval estimate. With  $n = 10$ ,  $\bar{x}_n = 3.22$ ,  $\hat{\sigma} = 1.17$ , a 95% confidence interval is given by  $(3.22 - (2.262)(1.17)/\sqrt{9}, 3.22 + (2.262)(1.17)/\sqrt{9}) = (2.34, 4.10)$   
(`display invt(9,.975)` gives you 2.262)

## Confidence Intervals for differences in means

Let  $X_1, \dots, X_n$  and  $Y_1, \dots, Y_m$  denote independent random samples from two distributions,  $N(\mu_1, \sigma^2)$  and  $N(\mu_2, \sigma^2)$ , with sample means denoted by  $\bar{X}$ ,  $\bar{Y}$  and sample variances by  $\hat{\sigma}_1^2$  and  $\hat{\sigma}_2^2$ .

We've established that:

- $\bar{X}$  and  $\bar{Y}$  are normally and independently distributed with means  $\mu_1$  and  $\mu_2$  and variances  $\frac{\sigma^2}{n}$  and  $\frac{\sigma^2}{m}$
- Using our results on the distribution of linear combinations of normally distributed variables, we know that  $\bar{X}_n - \bar{Y}_m$  is normally distributed with mean  $\mu_1 - \mu_2$  and variance  $\frac{\sigma^2}{n} + \frac{\sigma^2}{m}$ . The random variable  $\frac{(\bar{X}_n - \bar{Y}_m) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2}{m}}}$  has a standard normal distribution and will form the numerator of the  $T$  random variable that we are going to use.
- We also know that  $\frac{n\hat{\sigma}_1^2}{\sigma^2}$  and  $\frac{m\hat{\sigma}_2^2}{\sigma^2}$  have  $\chi^2$  distributions with  $(n-1)$  and  $(m-1)$  degrees of freedom respectively, so their sum  $(n\hat{\sigma}_1^2 + m\hat{\sigma}_2^2)/\sigma^2$  has a  $\chi^2$  distribution with  $(n+m-2)$  degrees of freedom and the random variable  $\sqrt{\frac{n\hat{\sigma}_1^2 + m\hat{\sigma}_2^2}{\sigma^2(n+m-2)}}$  can appear as the denominator of a random variable which has a  $t$ -distribution with  $(n+m-2)$  degrees of freedom.

## Confidence Intervals for differences in means..contd

- We have therefore established that  $X = \frac{(\bar{X}_n - \bar{Y}_m) - (\mu_1 - \mu_2)}{\sqrt{\frac{n\hat{\sigma}_1^2 + m\hat{\sigma}_2^2}{(n+m-2)} \left(\frac{1}{n} + \frac{1}{m}\right)}}$  has a t-distribution with  $(n + m - 2)$  degrees of freedom. To simplify notation, denote the denominator of the above expression by  $R$ .
- Given our samples,  $X_1, \dots, X_n$  and  $Y_1, \dots, Y_m$ , we can now construct confidence intervals for differences in the means of the corresponding populations,  $\mu_1 - \mu_2$ . We do this in the usual way:
  - Suppose we want a 95% confidence interval for the difference in the means, we find a number  $b$  such that, using the t-distribution with  $(n + m - 2)$  degrees of freedom,

$$\Pr(-b < X < b) = .95$$

- The random interval  $(\bar{X} - \bar{Y}) - bR, (\bar{X} - \bar{Y}) + bR$  will now contain the true difference in means with 95% probability.
- A confidence interval is now based on sample values,  $(\bar{x}_n - \bar{y}_m)$  and corresponding sample variances.
- Based on the CLT, we can use the same procedure even when our samples are not normal.