Topic 9: Sampling Distributions of Estimators

Rohini Somanathan

Course 003, 2018

Sampling distributions of estimators

Since our estimators are statistics (particular functions of random variables), their distribution can be derived from the joint distribution of $X_1 \ldots X_n$. It is called the sampling distribution because it is based on the joint distribution of the random sample.

Given a sampling distribution, we can

- make appropriate trade-offs between sample size and precision of our estimator since sampling distributions on sample size.
- 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 $\hat{\theta}$.

Application: 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:

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

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 θ. Suppose we want the probability that our estimate lies within $.1\theta$ for any level of θ 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 $\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$. We know that

$\bar{X}_n \sim N(\mu, \frac{\sigma^2}{n})$ **and** $\sum_{i=1}^{n}(\frac{X_i-\mu}{\sigma})^2 \sim \chi^2_n$ **(sum of squares of n standard normals)**

If we replace the population mean μ with the sample mean $\bar{X}_n$, the resulting sum of squares, has a χ^2_{n-1} distribution, which is independent of the distribution of $\bar{X}_n$. This is stated formally below:

Theorem: *If $X_1, \ldots X_n$ form a random sample from a normal distribution with mean μ and variance σ^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*

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

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

Note: This is only for normal samples.

Application: mean and variance estimates

We have a normal random sample and would like the M.L.E.s of the mean and standard deviation to be within one-fifth of a standard deviation of the respective parameters, μ and σ with some threshold probability.

- **Suppose we want to choose a sample size n such that $Pr(|\bar{X}_n - \mu| \leq \frac{1}{5}\sigma) \geq \frac{1}{2}$**
  - **If we use Chebyshev's inequality, we get this probability is greater than $\frac{25}{n}$, so setting this equal to $\frac{1}{2}$, we have $n = 50$**
  - **Using the exact distribution of $\bar{X}_n$, $Pr(|\bar{X}_n - \mu| \leq \frac{1}{5}\sigma) = Pr(\sqrt{n}\frac{|\bar{X}_n - \mu|}{\sigma} \leq \frac{1}{5}\sqrt{n})$**
  - **Since we now have a standard normal r.v., we know $Pr(Z > .68) = .25$, so we need the smallest n greater than $(.68 * 5)^2 = 11.6$, so $n = 12$ (Stata 14: invnormal(.75)=.6745)**
- **Now if we want to determine n so that $Pr[(|\bar{X}_n - \mu| \leq \frac{1}{5}\sigma$ and $(|\hat{\sigma}_n - \sigma| \leq \frac{1}{5}\sigma] \geq \frac{1}{2}$**
  - **By the previous theorem, $\bar{X}_n$ and $\hat{\sigma}_n$ are independent, so the LHS is the product $p_1 p_2 = Pr(|\bar{X}_n - \mu| \leq \frac{1}{5}\sigma)Pr(|\hat{\sigma}_n - \sigma| \leq \frac{1}{5}\sigma)$**
  - **$p_1 = Pr(|Z| \leq \frac{\sqrt{n}}{5}) = 1 - 2 * (1 - \Phi(\frac{\sqrt{n}}{5})$.**
  - **$p_2 = Pr(.8\sigma < \hat{\sigma}_n < 1.2\sigma) = Pr(.64n < \frac{n\hat{\sigma}_n^2}{\sigma^2} < 1.44n)$**
  - **Since $V = n\frac{\hat{\sigma}_n^2}{\sigma^2} \sim \chi^2_{n-1}$, we can search over values of n to find one that gives us a product of probabilities equal to $\frac{1}{2}$. For $n = 21$, $p_1 = .64$ $p_2 = .79$ so $p_1 p_2 = .5$.**
    display chi2(20, 30.24)-chi2(20, 13.44)
    (since 21*.64=13.44 and 21*1.44=30.24)

The t-distribution

Let $Z \sim N(0,1)$, let $Y \sim \chi^2_v$, 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:

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

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*

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

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

Implication: We cannot make statements about $|\bar{X}_n - \mu|$ using the normal distribution if σ^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 $\frac{(\overline{X}_n - \mu)}{\hat{\sigma}/\sqrt{n-1}} \sim t_{n-1}$

RESULT 2 As $n \to \infty$, $U \longrightarrow Z \sim N(0,1)$

To see why: U can be written as $\sqrt{\frac{n-1}{n}}\frac{\sqrt{n}(\overline{X}_n - \mu)}{\hat{\sigma}} \sim t_{n-1}$. As n gets large $\hat{\sigma}$ gets very close to σ 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 σ^2, let us see how we can obtain an interval estimate for μ, i.e. an interval which is likely to contain μ with a pre-specified probability.

Since $\frac{(\overline{X}_n-\mu)}{\sigma/\sqrt{n}} \sim N(0,1)$, $Pr\left(-2 < \frac{(\overline{X}_n-\mu)}{\sigma/\sqrt{n}} < 2\right) = .955$

But this event is equivalent to the events $-\frac{2\sigma}{\sqrt{n}} < \overline{X}_n - \mu < \frac{2\sigma}{\sqrt{n}}$ and $\overline{X}_n - \frac{2\sigma}{\sqrt{n}} < \mu < \overline{X}_n + \frac{2\sigma}{\sqrt{n}}$

With known σ, each of the random variables $\overline{X}_n - \frac{2\sigma}{\sqrt{n}}$ and $\overline{X}_n + \frac{2\sigma}{\sqrt{n}}$ are statistics. Therefore, we have derived a random interval within which the population parameter lies with probability .955, i.e.

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

Notice that there are many intervals for the same γ, this is the shortest one.

Now, given our sample, our statistics take particular values and the resulting interval either contains or does not contain μ. We can therefore no longer talk about the probability that it contains μ because the experiment has already been performed.

We say that $(\overline{x}_n - \frac{2\sigma}{\sqrt{n}} < \mu < \overline{x}_n + \frac{2\sigma}{\sqrt{n}})$ is a 95.5% confidence interval for μ. Alternatively, we may say that μ lies in the above interval with confidence γ or that the above interval is a confidence interval for μ with confidence coefficient γ

Confidence Intervals for means..examples

Example 1: $X_1, \ldots, X_n$ forms a random sample from a normal distribution with unknown μ and $\sigma^2 = 10$. $\bar{x}_n$ is found to be 7.164 with $n = 40$. An 80% confidence interval for the mean μ 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 (stata 14: display invnormal(.9))

Example 2: Let $\overline{X}$ denote the sample mean of a random sample of size 25 from a distribution with variance 100 and mean μ. In this case, $\frac{\sigma}{\sqrt{n}} = 2$ and, making use of the central limit theorem the following statement is approximately true:

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

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 σ^2. We know that $\frac{(\overline{X}_n - \mu)}{\hat{\sigma}/\sqrt{n-1}} \sim t_{n-1}$ 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 normal random samples.

$X_i \sim N(\mu_1, \sigma^2)$ and $Y_i \sim N(\mu_2, \sigma^2)$ respectively. Sample means and variances are $\bar{X}, \bar{Y}, \hat{\sigma}_1^2, \hat{\sigma}_2^2$.

We know (using previous results) that:

- $\bar{X}$ and $\bar{Y}$ are normally and independently distributed with means μ_1 and μ_2 and variances $\frac{\sigma^2}{n}$ and $\frac{\sigma^2}{m}$
- $(\bar{X}_n - \bar{Y}_m) \sim N(\mu_1 - \mu_2, \frac{\sigma^2}{n} + \frac{\sigma^2}{m})$ so $\frac{(\bar{X}_n - \bar{Y}_m) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2}{m}}} \sim N(0, 1)$
- $\frac{n\hat{\sigma}_1^2}{\sigma^2} \sim \chi^2_{n-1}$ and $\frac{m\hat{\sigma}_2^2}{\sigma^2} \sim \chi^2_{m-1}$, so their sum $(n\hat{\sigma}_1^2 + m\hat{\sigma}_2^2)/\sigma^2 \sim \chi^2_{n+m-2}$. Therefore

$$U = \frac{(\bar{X}_n - \bar{Y}_m) - (\mu_1 - \mu_2)}{\sqrt{\frac{n\hat{\sigma}_1^2 + m\hat{\sigma}_2^2}{(n+m-2)}\left(\frac{1}{n} + \frac{1}{m}\right)}} \sim t_{n+m-2}$$

  Denote the denominator of U by R.

Suppose we want a 95% confidence interval for the difference in the means:

- Using the above t-distribution, we find a number b for which $Pr\left(-b < X < b\right) = .95$
- 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.