

Econ 312: Introduction to Econometrics

Review of Probability III

Sang-Yeob Lee

February 16, 2010

Summary: The sampling Distribution of \bar{Y}

For Y_1, Y_2, \dots, Y_n i.i.d with $0 < \sigma^2 < \infty$,

- The exact (finite sample) sampling distribution of \bar{Y} has mean μ_Y and variance σ_Y^2/n
- Other than its mean and variance, the exact distribution of \bar{Y} is complicated and depends on the distribution of Y (the population distribution)
- When n is large, the sampling distribution simplifies:

-

$$\bar{Y}_n \rightarrow^p Y$$

(Law of large numbers)

-

$$\frac{\bar{Y} - E(Y)}{\sqrt{\text{Var}(\bar{Y})}}$$

is approximately $N(0,1)$ (CLT)

Why Use \bar{Y} to Estimate μ_Y ?

- \bar{Y} is unbiased: $E(\bar{Y}) = \mu_Y$
- \bar{Y} is consistent: $\bar{Y}_n \xrightarrow{p} Y$
- The best fits to the data in the sense that the average squared difference between the observation and \bar{Y} are the smallest of all possible estimators.
- \bar{Y} is the “least square” estimator of μ_Y : \bar{Y} solves, $\min_m \sum_{i=1}^n (Y_i - m)^2$ so, \bar{Y} minimizes the sum of squared “residuals”

$$\frac{d}{dm} \sum_{i=1}^n (Y_i - m)^2 = \sum_{i=1}^n \frac{d}{dm} (Y_i - m)^2$$

Set derivative to zero and denote optimal value of m by \hat{m} :
 $\sum_{i=1}^n Y_i = \sum_{i=1}^n \hat{m} = n\hat{m}$ or $\hat{m} = \frac{1}{n} \sum_{i=1}^n Y_i = \bar{Y}$.

Hypothesis Tests Concerning for the Population Mean

- Does the population mean of hourly earnings equal \$20?
- The hypothesis testing problem for the mean: make a provisional decision, based on the evidence at hand, whether the null hypothesis is true, or instead that some other alternative hypothesis is true. That is, test
 - $H_0 : E(Y) \leq \mu_0$ vs $H_1 : E(Y) > \mu_0$ (1-sided, > upper tail test)
 - $H_0 : E(Y) \geq \mu_0$ vs $H_1 : E(Y) < \mu_0$ (1-sided, < lower tail test)
 - $H_0 : E(Y) = \mu_0$ vs $H_1 : E(Y) \neq \mu_0$ (2-sided)
 - For example, the conjecture that, on average in the population, college graduates earn \$20/hours constitutes a null hypothesis about the population distribution of hourly earning. (Mathematically, $H_0 : E(Y) = 20$).

Some terminology for testing statistical hypothesis

- p-value: probability of drawing a statistic (e.g. \bar{Y}) at least as adverse to the null as the value actually computed with your data, assuming that the null hypothesis is true.
- For example, suppose that, in your sample, the average wage is \$ 22.24. The p-value is that probability of observing a value of \bar{Y} at least as different from \$20 as the observed value of \$22.24 by pure random sampling variation, assuming the null hypothesis is true.
 - If this p-value is small, say 0.5%, then it is very unlikely that this sample would have been drawn if the null hypothesis is true.
 - If this p-value is large, say 40%, then it is quite likely that this sample would have been drawn if the null hypothesis is true.

Calculating the p-value based on \bar{Y}

To compute the p-value, you need to know the sampling distribution of \bar{Y} , which is complicated if n is small.

If n is large, you can use the normal approximation (CLT)

- When σ is known
 - 2 sided: p-value=

$$\Pr\left(-\frac{\bar{Y} - \mu_0}{\sigma_Y / \sqrt{n}} > Z \mid H_0 : E(Y) = \mu_0\right) + \Pr\left(Z > \frac{\bar{Y} - \mu_0}{\sigma_Y / \sqrt{n}} \mid H_0 : E(Y) = \mu_0\right)$$

- 1 sided(>): p-value= $\Pr\left(Z > \frac{\bar{Y} - \mu_0}{\sigma_Y / \sqrt{n}} \mid H_0 : E(Y) \leq \mu_0\right)$
- 1 sided(<): p-value= $\Pr\left(Z < \frac{\bar{Y} - \mu_0}{\sigma_Y / \sqrt{n}} \mid H_0 : E(Y) \geq \mu_0\right)$
- In practice, σ_Y is unknown-it must be estimated.

Computing the p-value with σ_Y estimated:

- Estimator of the variance of Y :

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2$$

= “sample variance of Y ”

- 2 sided p-value

$$= \Pr\left(-\frac{\bar{Y} - \mu_0}{s_Y / \sqrt{n}} > t_{n-1} | H_0 : E(Y) = \mu_0\right) + \Pr\left(t_{n-1} < \frac{\bar{Y} - \mu_0}{s_Y / \sqrt{n}} | H_0 : E(Y) = \mu_0\right)$$

- Concept: The standard error of \bar{Y} is an estimator of the standard deviation of \bar{Y} That is $SE(\bar{Y}) = s_Y / \sqrt{n}$

What is the link between the p-value and the significance level?

- The significance level(α) of a test is a pre-specified probability of incorrectly rejecting the null, when the null is true.
- The significance level is pre-specified. For example, if the pre-specified significance level is 5%.
- 2 side
 - you reject the null hypothesis if $|t_{n-1}| > t_{0.5/2}$
 - equivalently, you reject if p-value ≤ 0.5
 - Often, it is better to communicate the p-value than simply whether a test rejects or nor- the p-value contains more information than the “yes/no” statement about whether the test rejects.

At this point, you might be wondering,

what happened to the t-table and the degrees of freedom?

- The t-statistic($= \frac{\bar{Y} - \mu_0}{SE(\bar{Y})}$) has the t-distribution with $n - 1$ degree of freedom.
- The critical value of t-distribution is tabulated in the back of all statistics books. Remember the recipe?
- 2 side
 - Compute t-statistic
 - Compute the degrees of freedom, which is $n - 1$
 - Look up the 5% critical value.
 - If the t-statistic exceed (in absolute value) this critical value, reject the null hypotheses.