Review of Probability and Statistics II

Zhenhao Gong

University of Connecticut

This course is designed to be:

1. Introductory
2. Leading by interesting questions and applications
3. Less math, useful, and fun!

Most important:

Feel free to ask any questions!

Enjoy! ☺

Goal

Reviews the **core** ideas of the theory of probability and statistics that are needed to understand regression analysis and econometrics

- The probability framework for statistical inference ✓
- Estimation
- Hypothesis Testing
- Confidence Intervals

- Randomness, random variable
- Population, population distribution (discrete and continuous)
- Moments: mean, variance, skewness and kurtosis
- Joint distribution and covariance, correlation
- Conditional distribution, conditional mean
- Sampling distribution

We use **the sampling distribution of** $\bar{Y}$ to do the statistical inference. So this is the **key** concept.

- $\bar{Y} = (Y_1 + Y_2 + \cdots + Y_n)/n$ is a random variable
- The distribution of $\bar{Y}$ over different possible samples of size n is called the sampling distribution of $\bar{Y}$
- The mean and variance of Y are the mean and variance of its sampling distribution, $E(\bar{Y})$ and $\text{Var}(\bar{Y})$

Question:
What are the differences from the population distribution of Y and its corresponding moments $E(Y)$ and $\text{Var}(Y)$?

Why we need to study the sampling distribution of $\bar{Y}$ and use it for estimation instead of studying the distribution of Y directly?

- For example, if we want to know how much the mean earnings differ for men and women within U.S., what can we do?
- Can we perform an exhaustive survey of the population of workers to find the population distribution of earnings?

The key insight of statistics is that one can learn about a population distribution by selecting a random sample from that population.

Example

Suppose Y takes on 0 or 1 (a Bernoulli random variable) with the probability distribution,

- $P(Y = 1) = p = 0.78$ and $P(Y = 0) = 1 - p = 0.22$
- $E(Y) = \mu_Y = p \times 1 + (1 - p) \times 0 = 0.78$
  $\text{Var}(Y) = E(Y - E(Y))^2 = \sigma_Y^2 = p(1 - p) = 0.1716$

The sampling distribution of $\bar{Y}$ depends on n. Consider $n = 2$, then the sampling distribution of $(\bar{Y})$ is

- $P(\bar{Y} = 0) = 0.22^2 = 0.0484$
- $P(\bar{Y} = 1/2) = 2 \times 0.22 \times 0.78 = 0.3432$
- $P(\bar{Y} = 1) = 0.78^2 = 0.6084$
- Distribution of $\bar{Y}$? Very complicated as n getting larger!!

Questions

Things we want to know about the sampling distribution of $\bar{Y}$, in order to use it to do estimation:

- What is $E(\bar{Y})$?
  If $E(\bar{Y}) = \mu_Y$(population mean) $= 0.78$, then $\bar{Y}$ is an unbiased estimator of μ_Y.
- What is $\text{Var}(\bar{Y})$? Does it also depends on n?
- Does $\bar{Y}$ become close to μ_Y when n is large?
- Is it true that $\bar{Y}$ is approximately normally distributed (appears bell shape) for large n?

The mean and variance of the sampling distribution of $\bar{Y}$ in general case, for Y_i i.i.d. from any distribution, not just Bernoulli:

- mean: $E(\bar{Y}) = E(\frac{1}{n}\sum_{i=1}^{n} Y_i) = \frac{1}{n}\sum_{i=1}^{n} E(Y_i) = \mu_Y$.
- variance:

$$\begin{aligned}\text{Var}(\bar{Y}) &= \text{Var}(\frac{1}{n}\sum_{i=1}^{n} Y_i) \\ &= \frac{1}{n^2}\text{Var}(\sum_{i=1}^{n} Y_i) = \frac{n\sigma_Y^2}{n^2} = \frac{\sigma_Y^2}{n}.\end{aligned}$$

- Implication? Unbiased estimator!

For small sample sizes, we can only know $E(\bar{Y})$ and $\text{Var}(\bar{Y})$. The sampling distribution of $\bar{Y}$ is complicated and depends on the distribution of Y, but if n is large, the sampling distribution of $\bar{Y}$ becomes simple!

1. As n increases, the distribution of $\bar{Y}$ becomes more tightly centered around μ_Y (**the Law of Large Numbers**)
2. Moreover, the distribution of $\bar{Y}$ (regardless the distribution of Y!) becomes normal (**the Central Limit Theorem**):
   - $\bar{Y}$ is approximately distributed $N(\mu_Y, \frac{\sigma_Y^2}{n})$
   - Standardized $\bar{Y}$: $Z = \sqrt{n}(\bar{Y} - \mu_Y)/\sigma_Y$ is approximately distributed $N(0, 1)$ (standard normal)

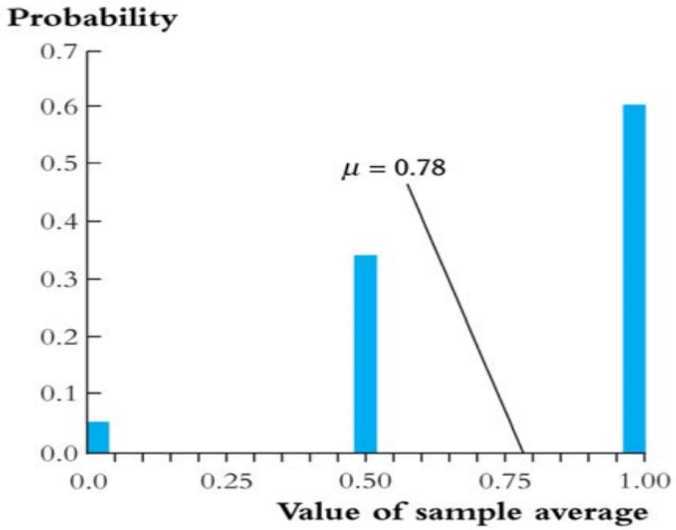
Recall example

Suppose Y takes on 0 or 1 (a Bernoulli random variable) with the probability distribution,

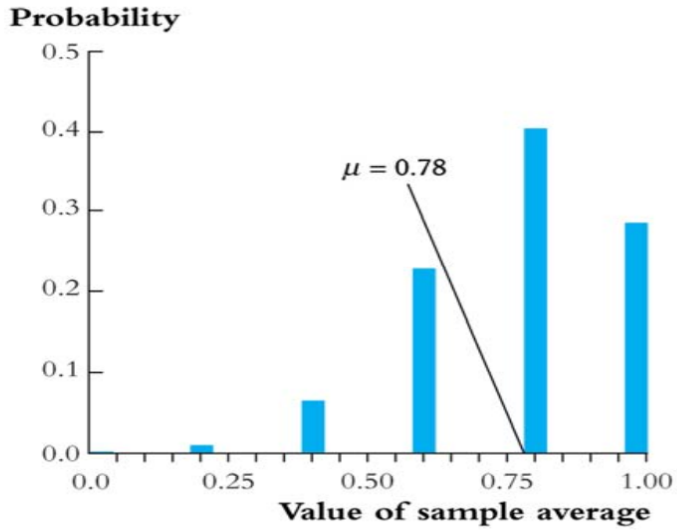
- $P(Y = 1) = p = 0.78$ and $P(Y = 0) = 1 - p = 0.22$
- $E(Y) = \mu_Y = p \times 1 + (1 - p) \times 0 = 0.78$

The sampling distribution of $\bar{Y}$ depends on n. Consider $n = 2$, then the sampling distribution of $(\bar{Y})$ is

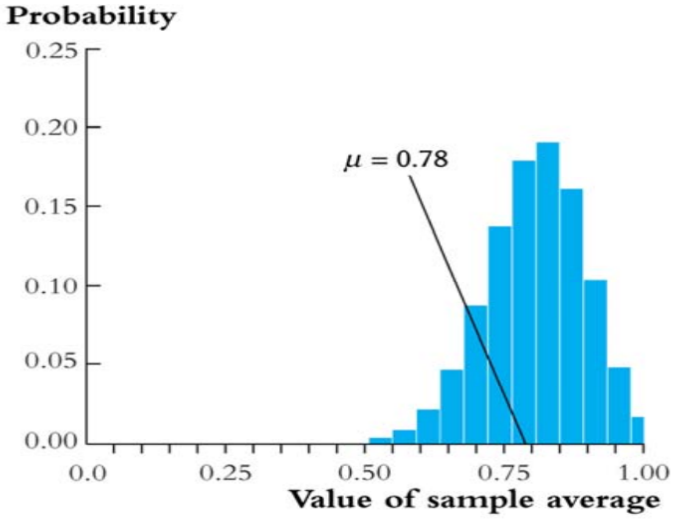
- $P(\bar{Y} = 0) = 0.22^2 = 0.0484$
- $P(\bar{Y} = 1/2) = 2 \times 0.22 \times 0.78 = 0.3432$
- $P(\bar{Y} = 1) = 0.78^2 = 0.6084$
- Distribution of $\bar{Y}$? Very complicated as n getting larger!!



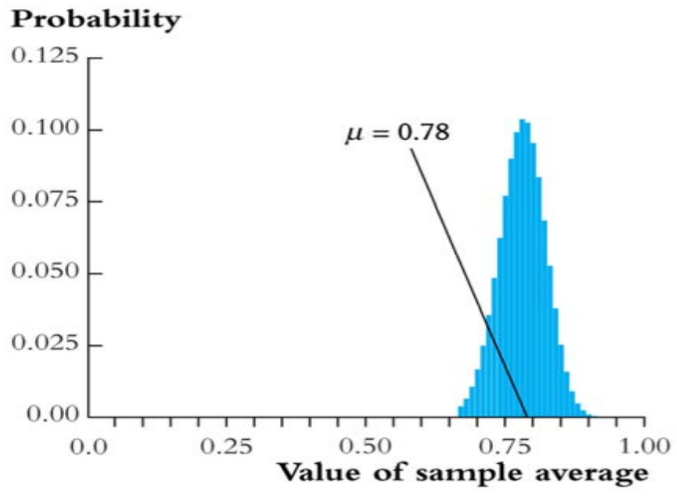
(a) $n = 2$



(b) $n = 5$

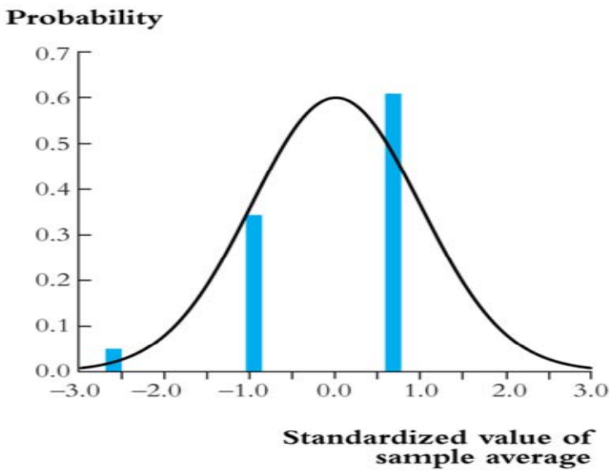


(c) $n = 25$

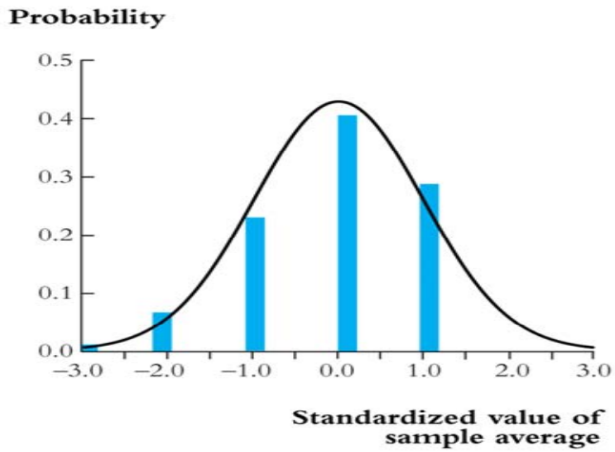


(d) $n = 100$

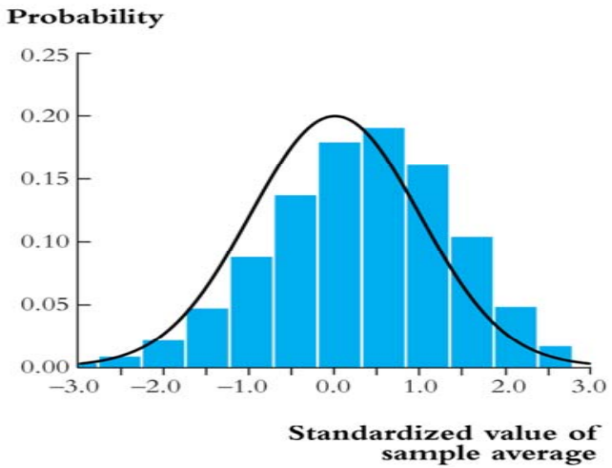
Same example: sampling distribution of $\dfrac{\bar{Y} - E(\bar{Y})}{\sqrt{\text{var}(\bar{Y})}}$



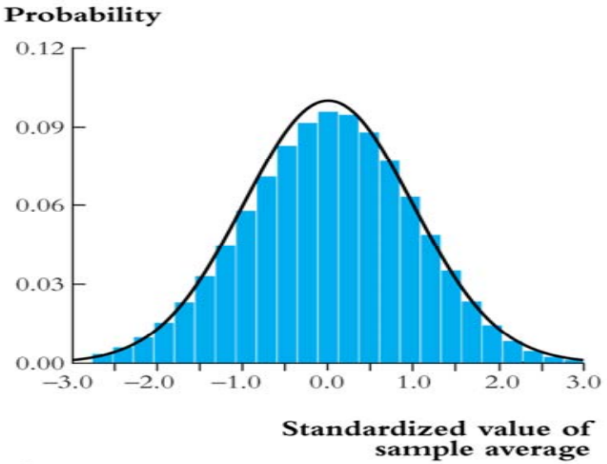
(a) $n = 2$



(b) $n = 5$

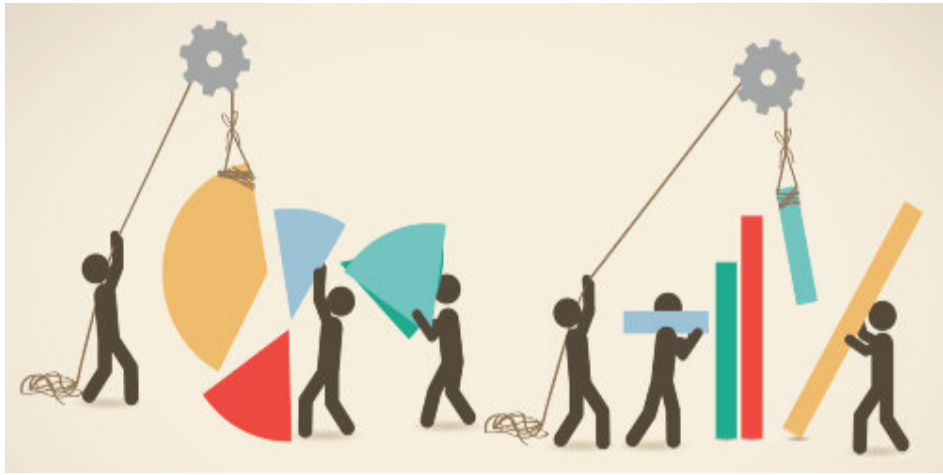


(c) $n = 25$



(d) $n = 100$

Review of Statistics



Goal

Reviews the **core** ideas of the theory of probability and statistics that are needed to understand regression analysis and econometrics

- The probability framework for statistical inference ✓
- Estimation ✓
- Hypothesis Testing
- Confidence Intervals

Hypothesis: yes/no question.

- Do the mean hourly earnings of recent U.S. college graduates equal 20 per hour?
- Are mean earnings the same for male and female college graduates?

Both questions concern about the population distribution of earnings and require evidence! The statistical challenge is to answer these questions based on a sample of evidence.

The **hypothesis testing** problem for the mean $E(Y)$:

- Make a **provisional** decision based on the evidence at hand whether a null hypothesis is true, or instead that some alternative hypothesis is true. That is, test

$$H_0 : E(Y) = \mu_{Y,0} \quad \text{vs.} \quad H_1 : E(Y) \neq \mu_{Y,0}$$

Example: the conjecture that, on average in the population, college graduates earn 20 per hour constitutes a null hypothesis about the population distribution of hourly earnings.

Remark: we can either rejecting the null hypothesis or failing to do so.

In any given sample, the sample average $\bar{Y}$ will rarely be exactly equal to the hypothesized value $\mu_{Y,0}$. Two reasons cause the difference:

- the null hypothesis is false
- the null hypothesis is true, but $\bar{Y}$ differs from $\mu_{Y,0}$ because of random sampling

It is impossible to distinguish between these two possibilities with certainty, but it is possible to do a probabilistic calculation that permits testing the null hypothesis in a way that accounts for sampling uncertainty.

P-value = the probability of drawing a value of $\bar{Y}$ that differs from $\mu_{Y,0}$ by at least as much as $\bar{Y}^{act}$, the value actually computed with your data, assuming that the null hypothesis is true.

Calculating the P-value based on $\bar{Y}$:

$$\text{P-value} = P_{H_0}[|\bar{Y} - \mu_{Y,0}| > |\bar{Y}^{act} - \mu_{Y,0}|],$$

where $\bar{Y}^{act}$ is the value of $\bar{Y}$ actually observed (nonrandom).

- To compute the p-value, we need the to know the sampling distribution of $\bar{Y}$, which is complicated if n is small.
- If n is large, we can use the normal approximation (CLT).

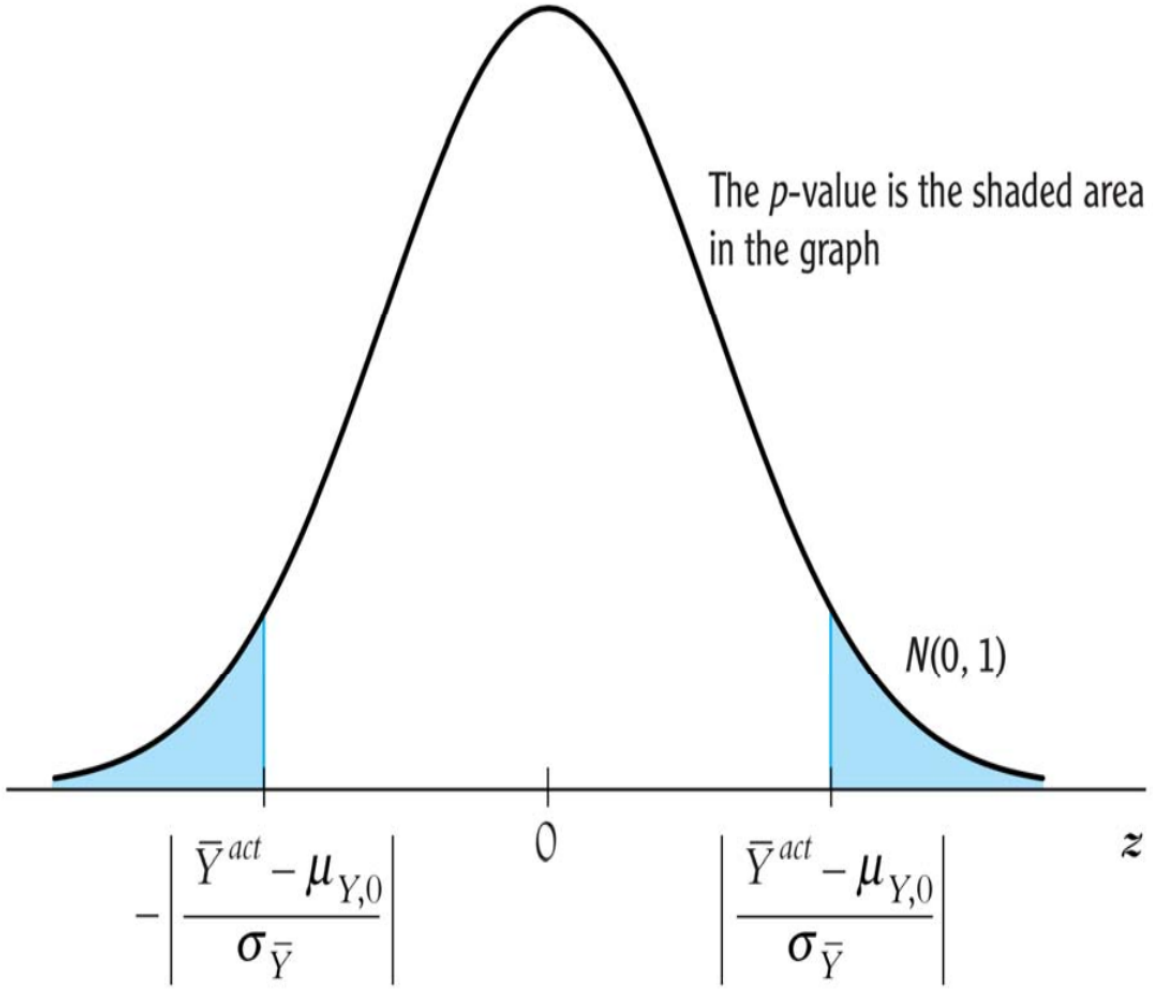
Compute P-value

When n is large, we know from CLT that the sampling distribution of $\bar{Y}$ is $N(\mu_{Y,0}, \sigma_Y^2/n)$. Let $\sigma_{\bar{Y}}^2 = \sigma_Y^2/n$, so $(\bar{Y} - \mu_{Y,0})/\sigma_{\bar{Y}}$ has a standard normal distribution. Hence, the P-value can be computed as

$$\begin{aligned}
\text{P-value} &= P_{H_0}[|\bar{Y} - \mu_{Y,0}| > |\bar{Y}^{act} - \mu_{Y,0}|] \\
&= P_{H_0}\left(\left|\frac{\bar{Y} - \mu_{Y,0}}{\sigma_{\bar{Y}}}\right| > \left|\frac{\bar{Y}^{act} - \mu_{Y,0}}{\sigma_{\bar{Y}}}\right|\right) \\
&= 2\Phi\left(-\left|\frac{\bar{Y}^{act} - \mu_{Y,0}}{\sigma_{\bar{Y}}}\right|\right),
\end{aligned}$$

where Φ is the standard normal cumulative distribution function.

Calculating the p-value with σ_Y known:



In practice, σ_Y is unknown and it can be estimated by the sample variance of Y:

$$s_Y^2 = \frac{1}{n-1}\sum_i^n (Y_i - \bar{Y})^2.$$

$$\Rightarrow \text{P-value} \approx P_{H_0}\left(\left|\frac{\bar{Y}-\mu_{Y,0}}{s_Y/\sqrt{n}}\right| > \left|\frac{\bar{Y}^{act}-\mu_{Y,0}}{s_Y/\sqrt{n}}\right|\right) = P_{H_0}(|t| > |t^{act}|),$$

where $t = \frac{\bar{Y}-\mu_{Y,0}}{s_Y/\sqrt{n}}$ the t-statistic, it approximate equal to standard normal distribution when n is large.

Example

Suppose we want to test the null hypothesis that the mean wage, $E(Y) = 20$ per hour using a sample of $n = 200$ recent college graduates.

- Step 1: compute the sample average wage $\bar{Y}^{act} = \$22.64$.
- Step 2: compute the sample standard deviation $s_Y = \$18.14$, so $s_Y/\sqrt{n} = \$18.14/\sqrt{200} = 1.28$.
- Step 3: compute the value of t-statistic $t^{act} = (22.64 - 20)/1.28 = 2.06$, so the p-value is $2\Phi(-2.06) = 0.039$ or 3.9%.

That is, assuming the null hypothesis is true, the probability of obtaining a sample average at least as different from the null as the one actually computed is 3.9%. Reject!

We can do the hypothesis test without computing the p-value. Instead, we can use a prespecified **significance level**. For example, if the prespecified significance level is 5%,

- we reject the null hypothesis if $|t^{act}| > 1.96$.
- equivalently, we reject if p-value ≤ 0.05.

This is often used in empirically studies. For example, we can test the null hypothesis under different significance levels. The most popular ones are $10\%, 5\%$ and 1%.

Goal

Reviews the **core** ideas of the theory of probability and statistics that are needed to understand regression analysis and econometrics

- The probability framework for statistical inference ✓
- Estimation ✓
- Hypothesis Testing ✓
- Confidence Intervals

Confidence Interval

A 95% confidence interval for μ_Y is an interval that contains the true value of μ_Y in 95% of repeated samples. It can always be constructed as the set of values of μ_Y not rejected by a hypothesis test with a 5% significance level:

$$\left\{\mu_Y : \left|\frac{\bar{Y}^{act} - \mu_Y}{s_Y/\sqrt{n}}\right| \leq 1.96\right\} = \{\mu_{Y:} - 1.96 \leq \frac{\bar{Y}^{act} - \mu_Y}{s_Y/\sqrt{n}} \leq 1.96\right\}$$

$$= \left\{\mu_Y \in \left(\bar{Y}^{act} - 1.96\frac{S_Y}{\sqrt{n}}, \bar{Y}^{act} + 1.96\frac{S_Y}{\sqrt{n}}\right)\right\}.$$

Example, given $\bar{Y}^{act} = \$22.64$ and $s_Y/\sqrt{n} = 1.28$, the 95% confidence interval for mean hourly earnings is:
$(22.64 - 1.96 \times 1.28, 22.64 + 1.96 \times 1.28) = (\$20.13, \$25.15)$.

Key question: Do districts with smaller classes have higher test scores? We can get the numerical evidence from the data set by following steps:

1. Compare average test scores in districts with low STRs to those with high STRs (**"estimation"**)
2. Test the "null" hypothesis that the mean test scores in the two types of districts are the same, against the "alternative" hypothesis that they differ (**"hypothesis testing"**)
3. Estimate an interval for the difference in the mean test scores, high v.s. low STR districts (**"confidence interval"**)

Initial data analysis: Compare districts with "small" (STR < 20) and "large" (STR ≥ 20) class sizes:

Class Size	Average score ($\bar{Y}$)	Standard deviation (s_Y)	n
Small	657.4	19.4	238
Large	650.0	17.9	182

1. ***Estimation*** of Δ = difference between group means
2. ***Test the hypothesis*** that $\Delta = 0$
3. Construct a ***confidence interval*** for Δ

The mean difference between two group samples:

$$\begin{aligned}\bar{Y}_{\mathrm{s}} - \bar{Y}_{\mathrm{l}} &= \frac{1}{n_{\mathrm{s}}} \sum_{i=1}^{n_{\mathrm{s}}} Y_i - \frac{1}{n_{\mathrm{l}}} \sum_{i=1}^{n_{\mathrm{l}}} Y_i \\ &= 657.4 - 650.0 \\ &= 7.4.\end{aligned}$$

Question:

- Is this result statistically significant? at what level?
- Is this a big enough difference to be important for school reform discussions, for parents, or for a school committee?

Difference-in-means test: compute the t-statistic,

$$t = \frac{\bar{Y}_s - \bar{Y}_l}{\sqrt{\frac{s_s^2}{n_s} + \frac{s_l^2}{n_l}}} = \frac{657.4 - 650.0}{\sqrt{\frac{11.4^2}{238} + \frac{17.92^2}{182}}} = \frac{7.4}{1.83} = 4.05 > 1.96,$$

so reject the null hypothesis (no difference between group means) at the 5% significance level. That is this result is statistically significant at 5% level.

Confidence interval

A 95% confidence interval for the difference between the means is,

$$\begin{aligned}\left(\bar{Y}_s - \bar{Y}_l\right) \pm 1.96 \times \sqrt{\frac{s_s^2}{n_s} + \frac{s_l^2}{n_l}} \\ = 7.4 \pm 1.96 \times 1.83 = (3.8, 11.0).\end{aligned}$$

Two equivalent statements:

1. The 95% confidence interval for Δ doesn't include 0;
2. The hypothesis that $\Delta = 0$ is rejected at the 5% level.

Simple Linear Regression

Read: Stock & Watson, Chapter 4.