Lecture 2: Statistical Review and Couterfactual Framework

Po-Chun Huang

National Chengchi University

March 12, 2021

Today

- Statistical Review
- Counterfactual Framework
- Randomized Controlled Trials

- Statistical Review

Statistical Review

Random Variable

- Random variable: one that takes on numerical values and has an outcome that is determined by and experiment
- Experiment: any procedure that can have, at least, in theory be infinitely repeated and has a well-defined sets of outcomes
- What characterizes a random variable is its probability distribution
- E.g. flip a coin

Statistical Review

Expectation

- The expectation of a random variable Y: $\mu_Y = E(Y)$
- Let the population size $= N$

$$E(Y) = \frac{1}{N}\sum_{i=1}^{N} y_i$$
$$= \sum_{i=1}^{J} y_i \cdot f(y_j)$$

- Expectations is weighted average of all possible values that the variable Y_i can take on, with weights given by the probability these values appear in the population.
- Expectation (and other probability attributes) is a parameter

Statistical Review

Sample Mean

- Use sample mean, $\bar{Y}$, to estimate the expectation
- $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} y_i$
- For a given population, there is only one $E(Y)$, while there are many $\bar{Y}$, depending on how we choose n and just who ends up in our sample

Statistical Review

Unbiasedness

- Given a random sample, $\bar{Y}$ is an unbiased for $E(Y)$
- $E(\bar{Y}) = E(Y)$
- In other words, if we were to draw infinitely many random samples, the average of the resulting $\bar{Y}$ across draws would be the underlying population mean.

Statistical Review

Consistency/Law of Large Number (LLN)

- Given a random sample, $\bar{Y}$ is a consistent estimator for $E(Y)$
- $\lim_{n\to\infty} P(|\bar{Y} - E(Y)| < \epsilon) = 1$ for any $\epsilon > 0$
- The LLN tells us that in large samples, the sample average is likely to be very close to the corresponding population mean.

Statistical Review

Variance and Standard Deviation

- In addition to expectations, we're interested in variability.
- Population variance: $V(Y) = \sigma_Y^2 = E[(Y - E[Y])^2]$
- Use sample variance to estimate population variance
- $S(Y)^2 = \frac{1}{n}\sum_{i=1}^{n}(Y - \bar{Y})^2$
- Standard deviation, σ_Y, is the square root of σ_Y^2

Statistical Review

Sampling Variance

- $V(\bar{Y}) = \frac{\sigma_Y^2}{n}$
- Variance of Y measures how much Y varies from person to person, while variance of $\bar{Y}$ measures how much Y varies from one sample of size n to the next
- Having a low variance is a good property of an estimator (statistic)
- What if $V(\bar{Y}) = 0$? You would always get the right answer

Statistical Review

Standard Error

- The standard deviation of a sample statistic like sample mean is called its standard error
- The standard error of the sample mean: $SE(\bar{Y}) = \frac{\sigma_Y}{\sqrt{n}}$
- Estimated standard error: $\hat{SE}(\bar{Y}) = \frac{S(Y)}{\sqrt{n}}$
- Every estimate discussed in this class (sample mean, diff. in means, OLS, IV, RD, DD, etc.) has an associated standard error.

Statistical Review

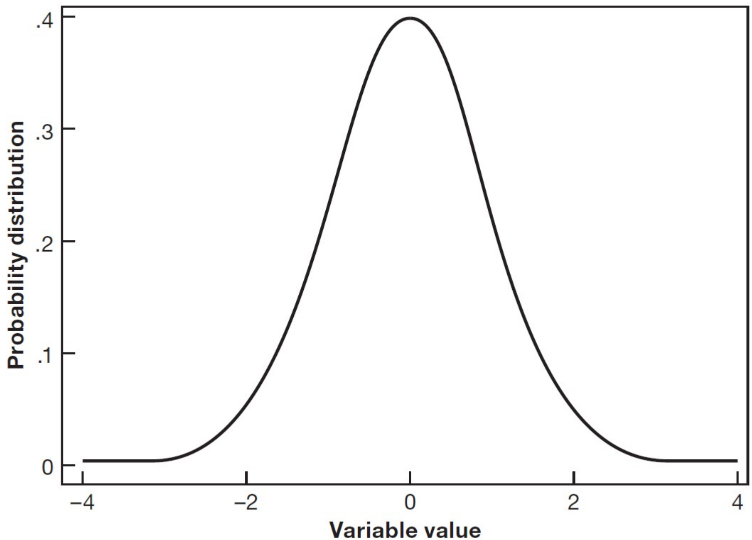
t Statistic

- Suppose we believe the population mean, $E(Y)$, takes on a particular value, μ
- This value constitutes a working hypothesis, a reference point that is often called the null hypothesis.
- A t-statistic for the sample mean under the working hypothesis that $E[Y] = \mu$:
$$t = \frac{\bar{Y} - \mu}{\hat{SE}(\bar{Y})}$$
- As long as the sample is large enough, the sampling distribution of t is approximately a standard normal distribution

Statistical Review

A Standard Normal Distribution

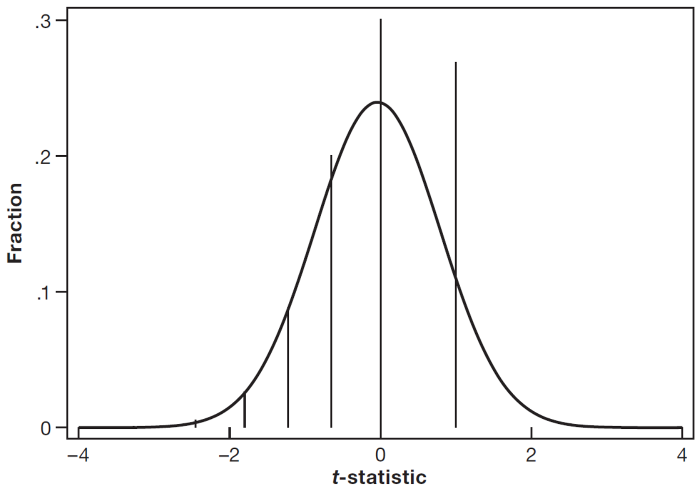
Figure 1.1
A standard normal distribution



- This property, which applies regardless of whether Y itself is normally distributed, is called the Central Limit Theorem (CLT).
- It implies that the (large-sample) distribution of a t-statistic is independent of the distribution of the underlying data used to calculate it.

- Suppose we measure health status with a dummy variable distinguishing healthy people from sick and that 20% of the population is sick.
- The distribution of this dummy variable has two spikes, one of height 0.8 at the value 1 and one of height 0.2 at the value 0.
- The CLT tells us that with enough data, the distribution of the t-statistic is smooth and bell-shaped even though the distribution of the underlying data has only two values.

FIGURE 1.2
The distribution of the t-statistic for the mean in a sample of size 10



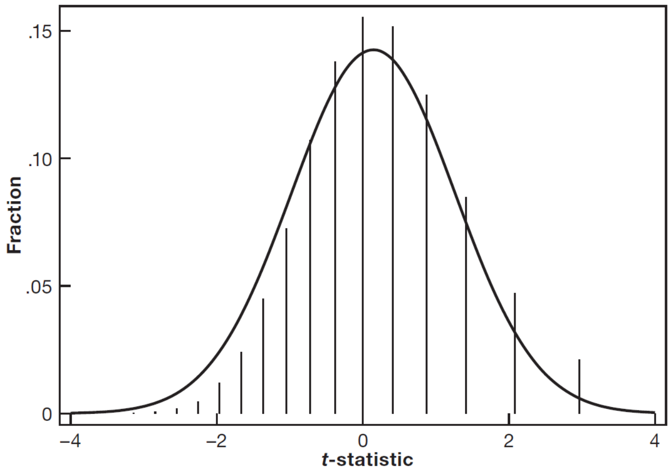
Note: This figure shows the distribution of the sample mean of a dummy variable that equals 1 with probability .8.

Statistical Review

Central Limit Theorem

Figure 1.3

The distribution of the t-statistic for the mean in a sample of size 40

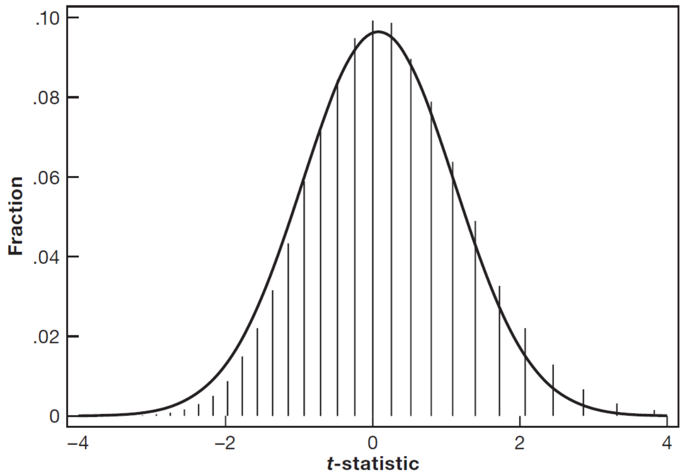


Note: This figure shows the distribution of the sample mean of a dummy variable that equals 1 with probability .8.

Statistical Review

Central Limit Theorem

FIGURE 1.4
The distribution of the t-statistic for the mean in a sample of size 100



Note: This figure shows the distribution of the sample mean of a dummy variable that equals 1 with probability .8.

Statistical Review

Hypothesis Testing

- With any standard normal variable, values larger than ± 2 are highly unlikely.
- In fact, realizations larger than 2 in absolute value appear only about 5% of the time.
- With a large sample, it's customary to judge any t-statistic larger than about 2 (in absolute value) as too unlikely to be consistent with the null hypothesis used to construct it.
- When the t-statistic exceeds 2 in absolute value, we say the sample mean is significantly different from μ and we reject the null.

Statistical Review

Confidence Interval

- Instead of checking whether the sample is consistent with a specific value of μ, we can construct the set of all values of μ that are consistent with the data.
- The set of such values is called a confidence interval for $E[Y]$
- When calculated in repeated samples, the following interval should contain $E[Y]$ about 95% of the time.

$$[\bar{Y} - 2\hat{SE}(\bar{Y}), \bar{Y} + 2\hat{SE}(\bar{Y})]$$

- This interval is therefore said to be a 95% confidence interval for the population mean.
- If the realization of the interval does not contain μ, we reject the null

Statistical Review

Extensions to Multiple Variables

- Up to this point, we have only worked with one random variable at a time
- In econometrics, we'll be interested in the relationship among many variables

- Joint density function:

$$f_{X,Y}(x, y) = P(X = x, Y = y)$$

- Marginal density function:

$$f_X(x) = P(X = x) = \sum_{y=-\infty}^{\infty} f_{X,Y}(x, y)$$

- Conditional density function:

$$f_{X|Y}(y|x) = \frac{f_{X,Y}(x, y)}{f_X(x)}$$

- Independence: independence means that two variables do not move together

$$X \amalg Y \text{ iff } f_{X,Y}(x, y) = f_X(x) \cdot f_Y(y)$$

- Covariance: a measure of how two variables move together

$$Cov(X, Y) = E[(X - E(X))(Y - E(Y))]$$

- Correlation: a scale-independent measure of how two variables move together

$$Corr(X, Y) = \frac{Cov(X, Y)}{s.d.(X) \cdot s.d.(Y)}$$

- Covariance and correlation measure linear dependence

Statistical Review

Extensions to Multiple Variables

- Conditional Expectation:

$$E(Y|X) = \sum_{y=-\infty}^{\infty} y \cdot f(y|x)$$

- e.g. "How do wages for the college educated compare to the wages for the high school educated?" is a question about conditional means
- If X and Y are independent, $E(Y|X) = E(Y)$
- $E[f(X)|X] = f(X)$

Statistical Review
Pairing Off

- Suppose you are interested in the earning difference between college graduates and high-school graduates
- Let μ_1 and μ_0 represent the earning for college graduates and high-school graduates
- You randomly draw n_1 college graduates and n_0 high-school graduates
- Estimate the earning difference using sample counterpart:

$$\bar{Y}_1 - \bar{Y}_0$$

Statistical Review

Pairing Off

- Sampling variance:

$$
\begin{aligned}
Var(\bar{Y}_1 - \bar{Y}_0) &= Var(\bar{Y}_1) + Var(\bar{Y}_0) \\
&= \frac{\sigma_Y^2}{n_1} + \frac{\sigma_Y^2}{n_0} \\
&= \sigma_Y^2 [\frac{1}{n_1} + \frac{1}{n_0}]
\end{aligned}
$$

- Standard error:

$$SE(\bar{Y}_1 - \bar{Y}_0) = \sigma_Y \sqrt{\frac{1}{n_1} + \frac{1}{n_0}}$$

- Estimated standard error:

$$S\hat{E(\bar{Y}_1 - \bar{Y}_0)} = S(Y) \sqrt{\frac{1}{n_1} + \frac{1}{n_0}}$$

Statistical Review

Pairing Off

- t-statistic for the difference in means:

$$t = \frac{\bar{Y}_1 - \bar{Y}_0 - 0}{\widehat{SE(\bar{Y}_1 - \bar{Y}_0)}}$$

- When the t-statistic is large enough to reject a difference of zero, we say the estimated difference is statistically significant.

- The confidence interval for a difference in means is the difference in sample means plus or minus two standard errors.

Statistical Review

Pairing Off

- Bear in mind that t-statistics and confidence intervals have little to say about whether findings are substantively large or small.
- The fact that an estimated difference is not significantly different from zero need not imply that the relationship under investigation is small or unimportant.
- Lack of statistical significance often reflects lack of statistical precision, that is, high sampling variance.

- Counterfactual Framework

Counterfactual Framework

- Suppose we are interested in the effects of college attendance (binary D) on earnings Y.
- Y_0: earnings if $D = 0$, Y_1: earnings if $D = 1$
- Everyone in population has values $\{D, Y_0, Y_1\}$
- Everyone has at least one of these missing (either Y_0 or Y_1, but not both).

Counterfactual Framework

- Treatment effect for unit $i = Y_{i1} - Y_{i0}$
- Average treatment effect $= E(Y_{i1} - Y_{i0})$
- Average treatment effect on the treated $= E(Y_{i1} - Y_{i0} | D = 1)$
- Average treatment effect on the untreated $= E(Y_{i1} - Y_{i0} | D = 0)$

Counterfactual Framework

- $ATE = ATT \cdot P(D=1) + ATUT \cdot P(D=0)$
- Observed earnings, $Y_i = Y_{i1} \cdot D_i + Y_{i0} \cdot (1 - D_i)$
- OLS estimate for the slope of linear regression of Y on D converge to?

Counterfactual Framework

- OLS slope estimate converges to expected difference in observed outcomes

$$E(Y|D=1) - E(Y|D=0)$$

$$= E(Y_1|D=1) - E(Y_0|D=0)$$

$$= E(Y_1 - Y_0|D=1) + E(Y_0|D=1) - E(Y_0|D=0)$$

$$= E(Y_1 - Y_0|D=0) + E(Y_1|D=1) - E(Y_1|D=0)$$

- Treatment Effect:

$$E(Y_1 - Y_0|D=1) \text{ and } E(Y_1 - Y_0|D=0)$$

- Selection Bias:

$$E(Y_0|D=1) - E(Y_0|D=0) \text{ and } E(Y_1|D=1) - E(Y_1|D=0)$$

Counterfactual Framework

- Observed difference in outcomes is a combination of treatment effect and a selection bias.

Solutions

- Random Assignment
- Regression and Matching
- Instrumental Variables
- Regression Discontinuity
- Panel data, Difference-in-Differences, and Synthetic Control

Randomized Controlled Trials

- If individuals are randomly assigned into treatment, then $D \amalg \{Y_0, Y_1\}$
- Random assignment eliminates selection bias:

$$E(Y_0|D=1) - E(Y_0|D=0) = E(Y_1|D=1) - E(Y_1|D=0) = 0$$

- When the treatment is randomly assigned, $ATT = ATUT = ATE$.