Econometrics Discussion Section 2

John Green

Spring 2024

Recap

What did we do in the first week?

- Review of probability and statistics (Stock & Watson chapters 2 and 3)
- Random variables and their first 4 moments
- Marginal and joint distributions
- LLN and CLT
- Properties of estimators: consistency, unbiasedness, normal approximation

Recap

What about in the second week?

- Estimator for σ^2
- One variable t-test
- Two variable t-test

Random variables

- A random variable is a *function* from a space of possible outcomes to (usually) the real numbers.
- Example:
  - Roll two dice and add up the numbers
  - Sample space is all the possible rolls of the two dice (how many?)
  - Outcome space is all the possible sums of those rolls (how many?)
  - The sum is a random variable

		FIRST DICE					
		1	**2**	**3**	**4**	**5**	**6**
SECOND DICE	**1**	2	3	4	5	6	7
	2	3	4	5	6	7	8
	3	4	5	6	7	8	9
	4	5	6	7	8	9	10
	5	6	7	8	9	10	11
	6	7	8	9	10	11	12

Mean

- The first moment of a random variable is its mean, aka average or expected value
- What is the mean value of the result from rolling one die?
- What is the mean value of the sum of rolling both dice?

Mean

- The first moment of a random variable is its mean, aka average, aka expected value
- What is the mean value of the result from rolling one die?
  - $E[X] = \frac{1}{6} \times 1 + \frac{1}{6} \times 2 + \frac{1}{6} \times 3 + \frac{1}{6} \times 4 + \frac{1}{6} \times 5 + \frac{1}{6} \times 6 = 3.5$
- What is the mean value of the sum of rolling both dice?
  - $E[X] = \frac{1}{36} \times 2 + \frac{2}{36} \times 3 + \frac{3}{36} \times 4 + \frac{4}{36} \times 5 + \frac{5}{36} \times 6 + \frac{6}{36} \times 7 + \frac{5}{36} \times 8 + \frac{4}{36} \times 9 + \frac{3}{36} \times 10 + \frac{2}{36} \times 11 + \frac{1}{36} \times 12 = 7$
  - (Could also write out using $\frac{1}{36}$ weight everywhere but easier to consolidate terms)

Variance

- The second moment of a random variable is its variance
    - Think of this as the mean distance from the mean: it measures the spread of our data
    - $Var(X) = E\left[(X - E[X])^2\right]$
    - Why do we need to square the spread?
- What is the variance of rolling one die?
- What is the variance of the sum of rolling both dice?

Variance

- The second moment of a random variable is its variance
  - Think of this as the mean distance from the mean: it measures the spread of our data
  - $Var(X) = E\left[(X - E[X])^2\right]$
  - Why do we need to square the spread?
- What is the variance of rolling one die?
  - $Var(X) = \frac{1}{6} \times (1 - 3.5)^2 + \frac{1}{6} \times (2 - 3.5)^2 + \cdots + \frac{1}{6} \times (6 - 3.5)^2 = \frac{35}{12} \approx 2.92$
- Similar idea for the sum of 2 dice.
- Often we will work with the *standard deviation* which is the square of the variance since the units are more interpretable.

Higher moments

- Third moment is *skewness* and tells us how symmetric our distribution is
  - What is the skewness of our example?
- Fourth moment is the *kurtosis* which measures the mass of the tails
  - Gives us an idea of the likelihood of large values

Covariance

- The covariance of two random variables tells us the strength of their *linear* (careful!) relationship
- If two random variables are independent, their covariance is 0
  - What is the covariance in our example?
- Often we will look at the *correlation* instead of the covariance, $\frac{cov(X,Y)}{(var(X)var(Y))^{1/2}}$
  - Unlike covariance, correlation is scaled between -1 and 1 and so is easily interpretable

Rules on joint distributions

- If two variables are independent, $P(X = x, Y = y) = P(X = x) * P(Y = y)$
- Conditional probability: $P(X = x|Y = y) = \frac{P(X=x,Y=y)}{P(Y=y)}$
    - This gives us the marginal distribution
    - What happens if the events are independent?
- Bayes' Rule: $P(X = x|Y = y) = \frac{P(Y=y|X=x)P(X=x)}{P(Y=y)}$
- Law of total probability: $P(X = x) = \sum_y P(X = x, Y = y)$

Second example

- New example:
  - Roll the first dice
  - If first roll ≥ 4 then we roll the second die and observe its value
  - If first roll ≤ 3 then the second value is simply set to 1
- What is the joint probability distribution?
- What is the covariance of the 2 rolls?
- What is the correlation?
- What is the marginal distribution of the two r.v.?

		FIRST DICE						
		1	2	3	4	5	6	Marginal:
SECOND DICE	1	0.17	0.17	0.17	0.03	0.03	0.03	0.58
	2	0	0	0	0.03	0.03	0.03	0.08
	3	0	0	0	0.03	0.03	0.03	0.08
	4	0	0	0	0.03	0.03	0.03	0.08
	5	0	0	0	0.03	0.03	0.03	0.08
	6	0	0	0	0.03	0.03	0.03	0.08
Marginal:		0.17	0.17	0.17	0.17	0.17	0.17	

Normal distribution

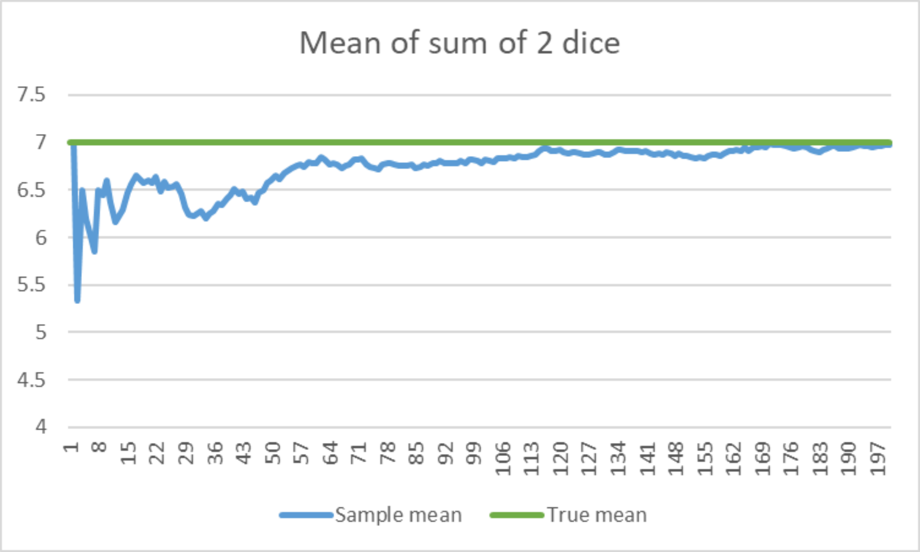
- Two parameters: mean μ and variance σ^2
- We can *standardize* a normal random variable X:
    - $Z = \frac{X-\mu}{\sigma}$ so $Z \sim N(0,1)$
    - Careful to divide by standard deviation and not by variance!
- This will be helpful in hypothesis testing because we can consider how unlikely a given value z is to be drawn from a N(0,1) distribution

Estimation

- We will often be interested in estimating the mean of a distribution
- Example: wait time in Brody Cafe
- Natural estimator is a sample mean:
  - Take a survey of people as they walk out of Brody and ask how long they waited
  - Average the responses
- Sample mean $\bar{Y}$ is a random variable since we taking a random sample and thus $\bar{Y}$ has a sampling distribution
  - What can we say about it?

Law of Large Numbers

- $E[\bar{Y}] = \mu_Y$
- This means that $\bar{Y}$ is an *unbiased* estimator of μ_Y
- As our sample size grows, the sample mean will converge to the true mean
- $var(\bar{Y}) = \frac{\sigma_Y^2}{n}$ so that the variance of our estimator is decreasing as our sample gets larger
- So as our sample size grows, the sample mean will converge to the true mean: $\bar{Y}$ is a *consistent* estimator of μ_Y (LLN)


Mean of sum of 2 dice
7.5
7
6.5
6
5.5
5
4.5
4
1
8
15
22
29
36
43
50
57
64
71
78
85
92
99
106
113
120
127
134
141
148
155
162
169
176
183
190
197
Sample mean
True mean

Central Limit Theorem

- Even better: as $n \to \infty$, $\bar{Y}$ becomes normal ie $\bar{Y} \sim N(\mu_Y, \frac{\sigma_Y^2}{n})$
- This means that we can use the normal distribution to make inferences about the sample mean
- To make it easy, we can standardize the sample mean: $Z = \frac{\bar{Y} - \mu_Y}{\sigma_Y / \sqrt{n}} \sim N(0, 1)$
- We will use the sample variance as an estimator for the population variance, just like we do for mean (but we will need to correct a small bias)

Hypothesis testing

- We can use the normal approximation to perform a *hypothesis test*
  - One-sided or two-sided
- Intuition: assuming the true mean is some value $\mu_{Y,0}$, how likely is it that we would observe the sample mean $\bar{Y}$?
  - If it is "very" unlikely, we will reject the null
  - If it is "reasonably likely" then we fail to reject the null
- p-value: probability of a test statistic at least as unlikely as the one you observe (under the null)

Some other terminology

- **Type-1 error**: reject a true null hypothesis
  - **Size** is probability of a type-1 error
- **Type-2 error**: fail to reject a false hypothesis
  - **Power** is probability of a type-2 error
- Which of these two mistakes is worse?
- What is the relationship between the size and the power?

Significance Level (α) and Power (1-β)

	H_0 is True	H_0 is False
Test Rejects H_0	α	1-β
Test Doesn't Reject H_0	1-α	β

Calculating p-value

- So, the **p-value** tells us how unlikely the null hypothesis is:
  - $P_{H_0}(|\frac{\bar{Y}-\mu_{Y,0}}{\sigma_Y/\sqrt{n}}| \geq |\frac{\bar{Y}_{data}-\mu_{Y,0}}{\sigma_Y/\sqrt{n}}|)$
  - RHS is just a number! LHS is RV with known distribution under null
  - CLT tells us this **t-statistic** is N(0,1) so probability in tails is easy to look up
- If observed t-statistic is very large or very small, then p-value is very small and we reject the null
  - If the null had been true, it is very unlikely that we would have observed the t-statistic

Using pre-set significance level

- Alternatively, we can decide that we want to do a test at a given significance α and find the value such that the sum in the tails (or tail) is α
- Eg if $\alpha = 0.05$, then our cutoff points are $(-1.96, 1.96)$
- If our t-statistic is outside of these bounds, we reject the null because it was very unlikely
- This is always less-informative than a p-value; but may be more rigorous to pre-set our rejection region
- We can also construct a confidence interval for μ_X based on our data, eg $\bar{X} \pm 1.96\frac{\sigma_X}{\sqrt{n}}$
  - This is the set of all possible μ_X values which would not be rejected by a two-sided t-test with $\alpha = 0.05$

Hypothesis about difference in means

- What if we want to perform a hypothesis on the difference in means between two random variables?
- Simple: combine them to one!
  - $X_3 = \bar{X}_1 - \bar{X}_2$
  - This is just a difference of two normal random variables so we can use the normal approximation, t-statistic is just $\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$
- What if we want to test hypothesis that the difference is 2?