ECON 340
Economics Research Methods

Div Bhagia

Midterm Review

Midterm Exam

- 1 hour 10 minutes, 20 points
- Closed book, can use a calculator
- Formula sheet and Normal Distribution table will be provided
- Study guide, formula sheet, and sample midterm is uploaded on the course website

Topics covered

- Summation notation
- Describing Data
    - Frequency distribution
    - Mean, median, variance, standard deviation
    - Add. topics: Percentiles, weighted mean, z-score
    - Covariance and correlation
- Random variables
    - Expected value and variance
    - Normal and standard normal distribution
    - Conditional expectation, uncorrelatedness and independence
- Sampling and Estimation
    - Sample mean distribution, CLT
    - Properties of a good estimator
    - Confidence intervals, hypothesis tests, p-values

Mean

Sample mean:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

The population mean is denoted by μ.

For grouped data:

$$\bar{X} = \frac{1}{n}\sum_{k=1}^{J} n_k X_k = \sum_{i=1}^{J} f_k X_k$$

Example: $X_i : 2, 2, -4, 2$

Variance

Population variance:

$$\sigma_X^2 = \frac{1}{N}\sum_{i=1}^{N}(X_i - \mu_X)^2$$

Sample variance:

$$S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Standard Deviation

$$\sigma_X = \sqrt{\sigma_X^2} \qquad S_X = \sqrt{S_X^2}$$

Variance: Grouped Data

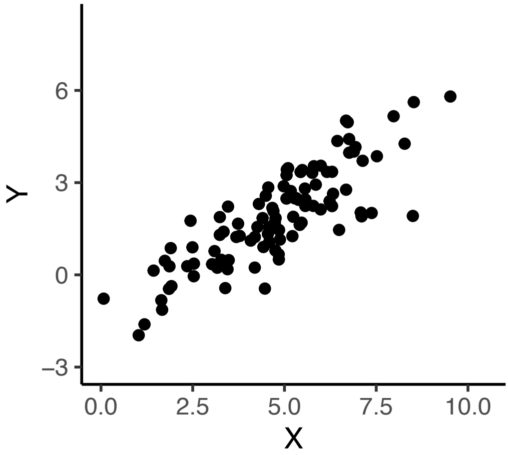
Alternatively, population variance:

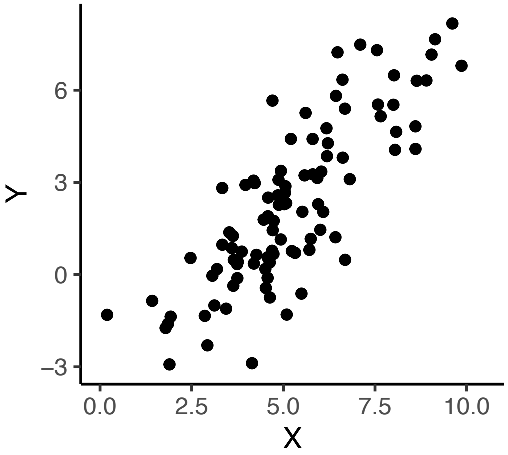
$$\sigma_X^2 = \sum_{k=1}^{J} f_k (X_k - \mu_X)^2$$

Sample variance:

$$S_X^2 = \frac{n}{n-1} \sum_{k=1}^{J} f_k (X_k - \bar{X})^2$$

The variance of Y is higher in which plot?





Weighted Mean

Weighted mean:

$$\bar{X}^{\omega} = \frac{\sum_{i=1}^{n} \omega_i X_i}{\sum_{i=1}^{n} \omega_i}$$

where ω_i is the weight of the i^{th} observation.

When weights sum up to 1 (i.e. $\sum_{i=1}^{n} \omega_i = 1$) we can simply write the weighted mean as:

$$\bar{X} = \sum_{i=1}^{n} \omega_i X_i$$

Example

- We want to estimate the average starting salary of students at a university that has only two majors
- Half of the students are *Business* majors, while the other half are *Engineering* majors
- Randomly select 100 Business students and 100 Engineering for a survey
- Response rate among Business students is 100%, while it 50% for engineering students

How can we use weighting to adjust for this?

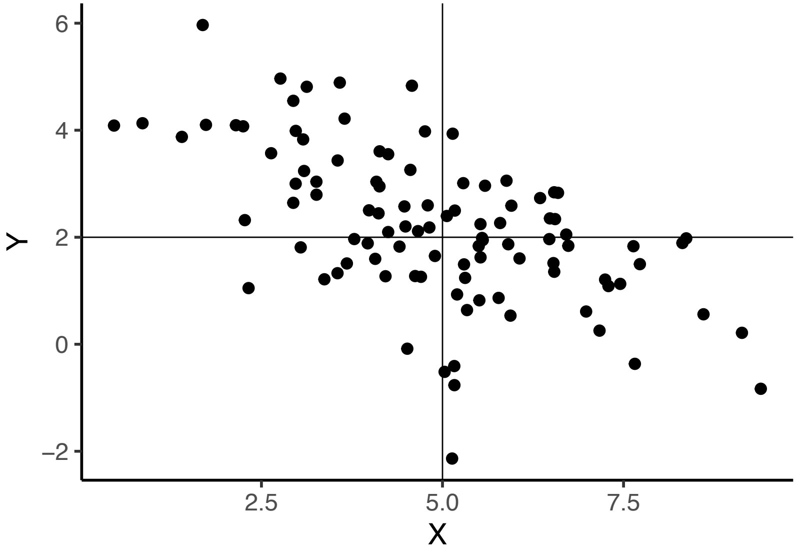
Covariance

$$\sigma_{XY} = \frac{1}{N}\sum_{i=1}^{N}(X_i - \mu_X)(Y_i - \mu_Y) \quad (Population)$$

$$S_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) \quad (Sample)$$

Why does the formula work?

Scatterplot



Correlation

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \quad (Population)$$

$$r_{XY} = \frac{S_{XY}}{S_X S_Y} \quad (Sample)$$

Why use correlation instead of covariance?

Correlation

- Measures the strength and direction of the linear relationship between two variables
- Bounded between -1 and 1
- If zero, there is no linear relationship. If 1 or -1 perfect linear relationship.
- If negative, when X is above (below) $\bar{X}$, Y tends to be above (below) $\bar{Y}$.
- If negative, when X is above (below) $\bar{X}$, Y tends to be below (above) $\bar{Y}$.

Correlation is not causation!

A positive correlation between job vacancies and immigration doesn't suggest that immigration $\rightarrow$ job creation. Why?

1. *Reverse causality:* jobs $\rightarrow$ immigration
2. *Other confounding factors:* government policies $\rightarrow$ jobs, government policies $\rightarrow$ immigration

Random Variables

A random variable is a variable that takes different values under different scenarios. Used for modeling uncertain outcomes.

Discrete RVs: Countable possible values

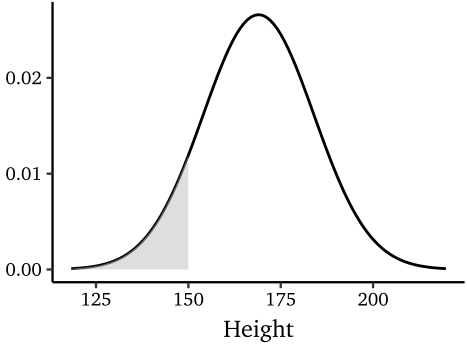
$$f(x) = Pr(X = x)$$

Continuous RVs: Any value in an interval

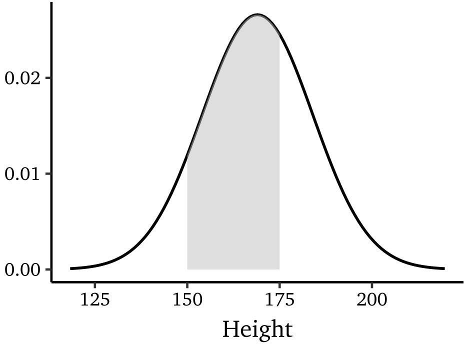
$$Pr(a \leqslant X \leqslant b) = \int_a^b f(x) \partial x$$

Normal Distribution

$Pr(X \leqslant 150)$



$Pr(150 < X < 175)$



Expectation & Variance

Discrete RV:

$$E(X) = \mu_X = \sum_x xf(x)$$

Continuous RV:

$$E(X) = \mu_X = \int_x xf(x)\partial x$$

Variance:

$$Var(X) = \sigma_X^2 = E[(X - \mu_X)^2]$$

Example

You estimate that the price of a stock will increase by 10% with a probability of 0.6 and decrease by 5% with a probability of 0.4. Calculate the expected return on the stock and its variance.

Covariance and Correlation

Covariance is a measure of the extent to which two random variables move together.

Let X and Y be a pair of random variables, then the *covariance* of X and Y is given by:

$$\sigma_{XY} = Cov(X,Y) = E[(X-\mu_X)(Y-\mu_Y)] = E(XY) - \mu_X\mu_Y$$

The *correlation* between X and Y is given by:

$$\rho_{XY} = corr(X,Y) = \frac{Cov(X,Y)}{\sigma_X\sigma_Y} \quad \text{where } -1 \leqslant \rho \leqslant 1$$

Conditional Distribution

The distribution of a random variable Y conditional on another random variable X taking on a specific value is called the conditional distribution of Y given X.

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

Example: Q2, Problem Set 3

Conditional Expectation

The *conditional expectation* of Y given X is the mean of the conditional distribution of Y given X.

$$E(Y|X = x) = \sum_{y} yPr(Y = y|X = x)$$

Example

X: Hours spent studying each week, Y: exam score

What does the following imply?

$$E(Y|X) = E(Y)$$

What if?

$$E(Y|X = 5) > E(Y|X = 1)$$

Z-score

Z-score is defined as:

$$Z = \frac{X - \mu}{\sigma}$$

Z-score tells us how many standard deviations any particular observation is away from the mean.

So if $X \sim N(\mu, \sigma^2/n)$, what is the distribution of Z?

Expectation and Variance of $\bar{X}$

Let $X_1, X_2, \ldots, X_n$ denote independent random draws (random sample) from a population with mean μ and variance σ^2.

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

The expectation and variance of the sample mean:

$$E(\bar{X}) = \mu \qquad Var(\bar{X}) = \sigma_{\bar{X}}^2 = \frac{\sigma^2}{n}$$

Distribution of the Sample Mean

What is the shape of the distribution for the sample mean?

It is normal when:

1. The underlying population is normal, or
2. The sample size is large, say $n \geqslant 100$, by the Central Limit Theorem

Confidence Intervals

- If $\bar{X} \sim N(\mu, \sigma^2_{\bar{X}})$, can create confidence intervals
- To create a 95% confidence interval, note:

$$Pr(-1.96 < Z < 1.96) = 0.95$$

- Since $Z = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}}$, we have

$$Pr(\mu - 1.96\sigma_{\bar{X}} < \bar{X} < \mu + 1.96\sigma_{\bar{X}}) = 0.95$$

- Which implies that:

$$Pr(\bar{X} - 1.96\sigma_{\bar{X}} < \mu < \bar{X} + 1.96\sigma_{\bar{X}}) = 0.95$$

Confidence Intervals

Let $z_{\alpha/2}$ be the z-value that leaves area $\alpha/2$ in the upper tail of the normal distribution.

Then $1 - \alpha$ confidence interval is given by

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

Note that, $\sigma_{\bar{X}} = \dfrac{\sigma}{\sqrt{n}}$

Hypothesis Testing: Recipe

1. Set up the null and alternative hypotheses:

$$H_0 : \mu = \mu_0 \qquad H_1 : \mu \neq \mu_0$$

2. Calculate test statistic:

$$z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$$

3. Reject the null if $|z| > z_{\alpha/2}$

p-Value

The p-value is defined as the probability of randomly drawing an outcome this surprising or even more surprising given the null hypothesis.

$$p = 2P(Z > |z|)$$

If p-value $< \alpha$, we reject the null.

When we don't know σ^2

Don't know the true population variance σ^2, use sample variance S^2.

The resulting test statistic:

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim t_{n-1}$$

follows a t distribution with $n - 1$ degrees of freedom.

In large samples, say $n \geqslant 100$, t is identical to standard normal so you can still refer to the standard normal table for critical values.

Example

A university wants to test whether online classes yield similar exam scores to in-person classes. Historically, the average exam score for traditional classes has been 75. A random sample of 100 exam scores from online classes reveals an average score of 71.5 with a standard deviation of 20.

- Test the hypothesis that scores from online classes are significantly different from in-person classes at a 10% level of significance.
- What is the p-value associated with your hypothesis test?
- What about testing this hypothesis at a 5% level of significance?
- Create a 95% confidence interval for average score from online classes.

Good luck!