Topic 1: Introduction to Forecasting

ECON 4753 — University of Arkansas

Prof. Kyle Butts

Fall 2024

Roadmap

Algebra Review

Linear Equations

Review of Statistics and Probability

Multiple Random Variables

Statistical Inference

Notation for Observations

When working with data, we often have many observations. To work with all observations at once, we use a notation: $x_1, x_2, \ldots, x_n$ where the subscript denotes which of the n observations we are referring to.

- This notation allows us to efficiently represent and manipulate large datasets

Summation Notation

When working with data, we often need to sum up the values of a variable for all observations.

- Instead of writing $x_1 + x_2 + \cdots + x_n$, we use the $\sum$ notation
- This notation is widely used in statistics, mathematics, and data analysis

The $\sum$ notation (Greek capital letter S for "Sum") allows us to concisely represent the sum of all observations.

Summation Notation

In general, the notation will look like this:

$$\sum_{i=1}^{n} x_i \tag{1}$$

- The subscript i is the iterator variable.
- The sum notation says: Start i at 1 ($i = 1$ part) and count up by one until you reach n.

Summation Notation

In general, the notation will look like this:

$$\sum_{i=1}^{n} x_i \tag{1}$$

- The subscript i is the iterator variable.
- The sum notation says: Start i at 1 ($i = 1$ part) and count up by one until you reach n.
- The $\sum$ term says "sum up all n terms iterated by i.
- x_i denotes what object to sum; in this case, sum the value of x for the i-th observation.

Example: Sum of Squares

Consider the following example:

Take $\sum_{i=1}^{5} i^2$. This says go from 1 to 5 and add the value of i^2.

This can be expanded as:

$$\sum_{i=1}^{5} i^2 = 1^2 + 2^2 + 3^2 + 4^2 + 5^2$$

Example: Sample Mean

Say you go out to the quad and start recording people's ages. You observe the following ten people: $\{19, 20, 32, 19, 22, 40, 28, 30, 19, 21\}$. To calculate the sample mean:

$$\frac{1}{10}\sum_{i=1}^{10} \text{Age}_i = \frac{1}{10}\left(19 + 20 + 32 + 19 + 22 + 40 + 28 + 30 + 19 + 21\right) = 25$$

Example: Sample Mean

Say you go out to the quad and start recording people's ages. You observe the following ten people: $\{19, 20, 32, 19, 22, 40, 28, 30, 19, 21\}$. To calculate the sample mean:

$$\frac{1}{10}\sum_{i=1}^{10}\text{Age}_i = \frac{1}{10}(19+20+32+19+22+40+28+30+19+21) = 25$$

In general, the mean is given by:

$$\frac{1}{n}\sum_{i=1}^{n} x_i$$

Properties of summation

It will be useful to look at a few special cases where we know what the sum will be.

For any constant (number) c,

$$\sum_{i=1}^{n} c = n * c$$

Properties of summation

It will be useful to look at a few special cases where we know what the sum will be.

For any constant (number) c,

$$\sum_{i=1}^{n} c = n * c$$

For any constant a,

$$\sum_{i=1}^{n} a x_i = a * \sum_{i=1}^{n} x_i$$

Properties of summation

Last, you can split up sums into parts:

$$\sum_{i=1}^{n}(x_i + y_i) = \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i$$

Properties of summation

Last, you can split up sums into parts:

$$\sum_{i=1}^{n}(x_i + y_i) = \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i$$

Putting them together, we have

$$\sum_{i=1}^{n}(a * x_i + b * y_i) = a * \sum_{i=1}^{n} x_i + b * \sum_{i=1}^{n} y_i$$

Application for our class

Variance

Define $\bar{x} = 1/n \sum_{i=1}^{n} x_i$ to be our sample mean. Let's work through the calculation of the *variance* of a variable.

The variance is defined as

$$\text{var}(x) \equiv \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Application for our class

Variance

$$\begin{aligned}\operatorname{var}(x) &\equiv \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \\ &= \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i^2 - 2 * x_i * \bar{x} + \bar{x}^2\right)\end{aligned}$$

Application for our class

Variance

$$\begin{aligned}
\mathrm{var}(x) &\equiv \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2 \\
&= \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i^2-2*x_i*\bar{x}+\bar{x}^2\right) \\
&= \frac{1}{n-1}\left(\sum_{i=1}^{n}x_i^2-\sum_{i=1}^{n}2*x_i*\bar{x}+\sum_{i=1}^{n}\bar{x}^2\right)
\end{aligned}$$

Application for our class

Variance

$$\begin{aligned}\operatorname{var}(X) &= \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} 2 * x_i * \bar{x} + \sum_{i=1}^{n} \bar{x}^2\right)\\ &= \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - 2 * \bar{x}\sum_{i=1}^{n} x_i + n * \bar{x}^2\right)\end{aligned}$$

Application for our class

Variance

$$
\begin{aligned}
\text{var}(X) &= \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} 2 * x_i * \bar{x} + \sum_{i=1}^{n} \bar{x}^2\right) \\
&= \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - 2 * \bar{x} \sum_{i=1}^{n} x_i + n * \bar{x}^2\right) \\
&= \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - 2 * \bar{x} * n * \bar{x} + n * \bar{x}^2\right)
\end{aligned}
$$

Application for our class

Variance

$$\begin{aligned}
\mathrm{var}(X) &= \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} 2 * x_i * \bar{x} + \sum_{i=1}^{n} \bar{x}^2\right) \\
&= \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - 2 * \bar{x} \sum_{i=1}^{n} x_i + n * \bar{x}^2\right) \\
&= \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - 2 * \bar{x} * n * \bar{x} + n * \bar{x}^2\right) \\
&= \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 + n * \bar{x}^2\right)
\end{aligned}$$

Review Questions

1. Evaluate the following:
   1.1 $\sum_{i=1}^{4}(i-2)$
   1.2 $\sum_{i=1}^{4}(i-1)^2$
   1.3 $\sum_{j=5}^{10} i$

Review Questions

1. Evaluate the following:
    1.1 $\sum_{i=1}^{4}(i-2)$
    1.2 $\sum_{i=1}^{4}(i-1)^2$
    1.3 $\sum_{j=5}^{10} i$

2. Write the sample mean of the variable Height (in.) in summation notation. What is the sample mean of the following set of observations $\{68, 66, 67, 70, 65, 66\}$?

Roadmap

Algebra Review

Linear Equations

Review of Statistics and Probability

Multiple Random Variables

Statistical Inference

Predicting Variable Y using Variable X

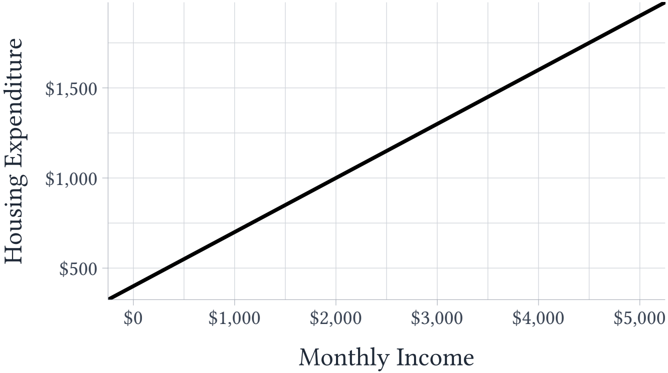
We are interested in predicting variable y using variable x. The relationship can be written as $y = f(x)$ for some function f.

In this section, we will focus on the linear equation, which takes the form

$$y = \beta_0 + \beta_1 * x$$

Regression Line Example

On average, we can calculate the average housing expenditure based on monthly income: Housing Expenditure $=$ $400 + 0.35 *$ Monthly Income:



Predictions

The first thing we can do is plug in a specific income, say $\$2500$, and determine the housing expenditures:

$$\text{Housing Expenditure} = 400 + 0.35 * 2500 = \$1275$$

- When our line is a *fitted model*, we call these predictions *fitted values* as you take the value of your explanatory variables and use the model fit to predict the value of the outcome variable

Marginal Effects

Also, we can think about changing someone's income and seeing how the outcome variable changes. Say you change from x_1 to x_2, how does y change in response?

$$y_2 - y_1 = (\beta_0 + \beta_1 * x_2) - (\beta_0 + \beta_1 * x_1)$$

$$= \beta_1(x_2 - x_1).$$

Marginal Effects

Also, we can think about changing someone's income and seeing how the outcome variable changes. Say you change from x_1 to x_2, how does y change in response?

$$\begin{aligned} y_2 - y_1 &= (\beta_0 + \beta_1 * x_2) - (\beta_0 + \beta_1 * x_1) \\ &= \beta_1 (x_2 - x_1). \end{aligned}$$

More succinctly, we can write $\Delta y = \beta_1 * \Delta x$

- Δ (greek for D for "difference") takes the difference between the new and old values of the variable

Marginal Effects

$$y_2 - y_1 = \beta_1(x_2 - x_1).$$

Notice the slope plays an important role here. It tells us when you change x by a certain amount, then y changes by that amount *scaled* by β_1.

- So, if I increase x by one unit, then y changes by β_1 units

Marginal Effects

$$y_2 - y_1 = \beta_1(x_2 - x_1).$$

Notice the slope plays an important role here. It tells us when you change x by a certain amount, then y changes by that amount *scaled* by β_1.

- So, if I increase x by one unit, then y changes by β_1 units

In our housing example, Housing Expenditure $= 400 + 0.35 *$ Monthly Income.

- If I increase my income by 500, then I change my housing expenditure by $0.35 * 500 = 175$ dollars.

Multiple Variables

Say we have the following $y = \beta_0 + \beta_1 x_1 + \beta_2 + x_2$, where x_1 and x_2 are two explanatory variables. Using simlar math above, show that

$$\Delta y = \beta_1 \Delta x_1 + \beta_2 \Delta x_2$$

Marginal Effects

$$\Delta y = \beta_1 \Delta x_1 + \beta_2 \Delta x_2$$

How does y change when you change x_1 while holding x_2 equal?

- We have $\Delta x_2 = 0$, so $\Delta y = \beta_1 \Delta x_1$ just like before.
- Called ‘marginal effect’ because we only change one variable, holding the rest fixed.

Example: Predicting Quantity Demanded

Consider the example of quantity demanded Q being a linear function of product price p and disposable income I:

$$Q = 120 - 9.8 * p + 0.03 * I$$

Say disposable income stays fixed at $\$900$ but the price increases from 8 to 9 dollars. How will the quantity demanded change?

Example: Predicting Quantity Demanded

Consider the example of quantity demanded Q being a linear function of product price p and disposable income I:

$$Q = 120 - 9.8 * p + 0.03 * I$$

Say disposable income stays fixed at $\$900$ but the price increases from 8 to 9 dollars. How will the quantity demanded change?

$$\Delta Q = -9.8 * \Delta p + 0.03 * \Delta I$$

$$= -9.8 * 1 + 0.03 * 0 = -9.8 \text{ units}$$

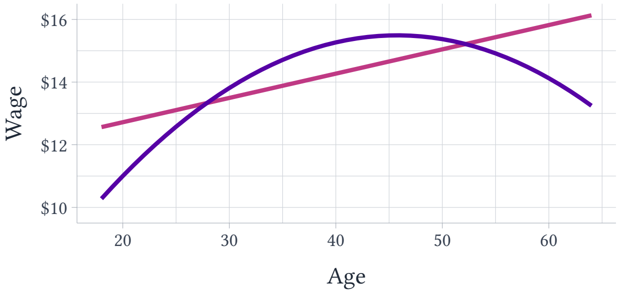
Quadratic model

Now say you model the relationship between x and y with a quadratic function

$$y = \beta_0 + \beta_1 x + \beta_2 x^2$$

Adding polynomial terms allows the relationship between x and y to not be linear

Example of Quadratic Function



The quadratic equation better represents the facts that (i) workers are promoted more often at younger ages (quicker wage growth) and (ii) earnings peek around age 45 for workers.

Marginal Effects

The equation for wages is given by

$$\text{Wage} = -5.5 + 1 * \text{Age} - 0.01 * \text{Age}^2$$

Given this equation, how do wages change as a person ages?

Marginal Effects

The equation for wages is given by

$$\text{Wage} = -5.5 + 1 * \text{Age} - 0.01 * \text{Age}^2$$

Given this equation, how do wages change as a person ages?

Using similar math from before, we can show:

$$\Delta\text{Wage} = (1 - 2 * 0.01 * \text{Age})\, \Delta\text{Age}$$

Marginal Effects

$$\Delta\text{Wage} = (1 - 2 * 0.01 * \text{Age})\,\Delta\text{Age}$$

How much wages change depend on what the worker's current age is

- Moving from 25 to 30, wages increase by $(1 - 2 * 0.01 * 25) * (30 - 25) = +\$2.50$

Marginal Effects

$$\Delta \text{Wage} = (1 - 2 * 0.01 * \text{Age})\, \Delta \text{Age}$$

How much wages change depend on what the worker’s current age is

- Moving from 25 to 30, wages increase by $(1 - 2 * 0.01 * 25) * (30 - 25) = +\$2.50$
- Moving from 55 to 60, wages decrease by $(1 - 2 * 0.01 * 55) * (60 - 55) = -\$0.50.$

Marginal Effects

More generally, the formula is given by

$$\Delta y = \underbrace{(\beta_1 + 2 * \beta_2 * x)}_{\text{depends on starting } x} * \Delta x$$

$\log$ transformation

Sometimes it is beneficial to log either the outcome and/or explanatory variables.

$\log$-transformations change how we interpret the regression line, namely by changing from ‘unit’ changes to ‘percent’ changes.

$\log$ transformation

Sometimes it is beneficial to log either the outcome and/or explanatory variables. $\log$-transformations change how we interpret the regression line, namely by changing from ‘unit’ changes to ‘percent’ changes.

E.g. I could model $\log(\text{Wages})$ as a linear regression of age

- My interpretation would be that increasing age by 1 year would increase wages by $\beta_1 * 100$ percent.

Summary of log-transformed linear equations

There are four models to consider

Model	**Interpretation**
$y = \beta_0 + \beta_1 x$	A 1 unit increase in x increases y by β_1 units
$\log(y) = \beta_0 + \beta_1 x$	A 1 unit increase in x increases y by $\beta_1 * 100$ percent
$y = \beta_0 + \beta_1 \log(x)$	A 1 percent increase in x increases y by β_1 units
$\log(y) = \beta_0 + \beta_1 \log(x)$	A 1 percent increase in x increases y by $\beta_1 * 100$ percent

We derive this in the review notes.

Roadmap

Algebra Review

Linear Equations

Review of Statistics and Probability

Multiple Random Variables

Statistical Inference

Key Concepts in Probability

Two key concepts in probability:

- Experiment: The source of randomness in the world (e.g. flip a coin, roll a dice, play a basketball game)
- Random Variable: Assigns numerical values determined by an experiment (e.g. value of a die roll, outcome of a game)

Notation for Random Variables

Notation for random variables:

- Upper case letters (W, X, Y, Z) denote random variables
- Lower case letters (w, x, y, z) denote particular values *realized* by the random variables

Random Variables in Coin Flipping

Example: Coin Flipping Experiment

X denotes the random variable counting the number of heads in 10 flips

- X represents the *process of running the experiment*, *not a particular value*
- X can take on values in the set $0, 1, 2, \ldots, 10$
- For a particular trial, X takes on a specific value, e.g. $x = 6$

Discrete random variables

A random variable is a discrete random variable when it takes on a finite amount of *distinct* values. E.g.:

- Number of heads out of 10 coin flips
- Student's final letter grade
- Whether a candidate wins an election (= 1) or not (= 0)

Probability Density Function

The probability density function (PDF) assigns probabilities for X obtaining any particular value:

Formally, for each possible value x_j of X, the PDF is defined as:

$$f_X(x_j) = \mathbb{P}(X = x_j) = p_j$$

Example pdf

Say you flip a single coin and assign X to equal 1 if the coin lands on heads and 0 if it lands on tails. The pdf for this random variable is $f(1) = 1/2$ and $f(0) = 1/2$.

Example pdf

Say you flip a single coin 3 times and record the number of heads in a random variable Y. The PDF is given by:

$$P_Y(0) = 1/8$$

$$P_Y(1) = 3/8$$

$$P_Y(2) = 3/8$$

$$P_Y(3) = 1/8$$

Properties of the Probability Density Function (PDF)

Two Rules for the PDF:

1. $0 \leq f_X(x_j) \leq 1$ for all values of x_j (probability of each value is between 0 and 1)

Properties of the Probability Density Function (PDF)

Two Rules for the PDF:

1. $0 \leq f_X(x_j) \leq 1$ for all values of x_j (probability of each value is between 0 and 1)
2. $\sum_j f_X(x_j) = 1$ (sum of probabilities over all possible values equals 1)

Cumulative Distribution Function (CDF)

The cumulative distribution function (CDF) is similar to the PDF but asks about the probability that X takes a value less than or equal to some x:

$$F_X(x) = \mathbb{P}(X \leq x)$$

Example CDF

Take a random variable, Y with pdf of $p_1 = \mathbb{P}(Y = 1) = 0.2$, $p_2 = \mathbb{P}(Y = 2) = 0.5$, and $p_3 = \mathbb{P}(Y = 4) = 0.3$. What is the CDF?

Example CDF

Take a random variable, Y with pdf of $p_1 = \mathbb{P}(Y = 1) = 0.2$, $p_2 = \mathbb{P}(Y = 2) = 0.5$, and $p_3 = \mathbb{P}(Y = 4) = 0.3$. What is the CDF?

$$F_y(y) = \begin{cases} 0.0 & \text{when } y < 1, \\ 0.2 & \text{when } 1 \leq y < 2, \\ 0.7 & \text{when } 2 \leq y < 4, \\ 1.0 & \text{when } y \geq 4, \end{cases}$$

Properties of the Cumulative Density Function (CDF)

Two Rules for the CDF:

1. $0 \leq F_X(x) \leq 1$

Properties of the Cumulative Density Function (CDF)

Two Rules for the CDF:

1. $0 \leq F_X(x) \leq 1$
2. $F_X(x)$ is (weakly) increasing in x

   $\rightarrow$ $P(X \leq x)$ can not be larger than $P(X \leq x +$ a little$)$

Probability Density Function and Statistics

The probability density function (pdf) provides a complete picture about a random variable, but it can often be too much information.

- Plotting the pdf as a histogram can still be overwhelming.

Probability Density Function and Statistics

The probability density function (pdf) provides a complete picture about a random variable, but it can often be too much information.

- Plotting the pdf as a histogram can still be overwhelming.

To summarize a random variable, we use statistics of the data

Expectation (Mean)

The most common statistic we use is the mean, also known as the expectation.

- It answers the question: ‘what is the average value of some random variable X?’

The expectation takes the average of the values that X can take, weighted by their probabilities:

$$\mathbb{E}(X) \equiv \sum_{j=1}^{n} x_j \, \mathbb{P}(X = x_j) = \sum_{j=1} x_j p_j$$

Example of Expectation

Consider a random variable Y with the following pdf:

- $\mathbb{P}(Y = 1) = 0.2$
- $\mathbb{P}(Y = 2) = 0.5$
- $\mathbb{P}(Y = 4) = 0.3$

The expectation of Y is:

$$\mathbb{E}(Y) = 1 \cdot 0.2 + 2 \cdot 0.5 + 4 \cdot 0.3 = 2.4$$

Properties of Expectations

The following properties of expectations can be derived from the definition of sums:

1. For any constant c, $\mathbb{E}(c) = c$ (no randomness)
2. For any constant a, $\mathbb{E}(aX) = a\,\mathbb{E}(X)$
3. For any constants a and b, $\mathbb{E}(aX + bY) = a\,\mathbb{E}(X) + b\,\mathbb{E}(Y)$

Transformations of Random Variables

Consider a function g that transforms a random variable X into a new random variable Y, such that $Y = g(X)$

The average value of Y can be calculated using the pdf of X:

$$\mathbb{E}(Y) = \mathbb{E}(g(X)) = \sum_{j=1}^{n} p_j g(x_j)$$

Variance

The other most common statistic is the variance of a random variable.

While the expectation tells you about the ‘average’ value, the variance tells you about the variability of the random variable:

- The variance measures how much the random variable X moves around its mean
- A large variance means the variable moves around a lot

Variance Formula

The variance of a random variable X is given by:

$$\operatorname{var}(X) \equiv \mathbb{E}\Big((X - \mathbb{E}(X))^2\Big) = \sum_j p_j (x_j - \mathbb{E}(X))^2,$$

where the last equality comes from the $g(X)$ rule

Intuition behind Variance

$$\text{var}\,(X) \equiv \mathbb{E}\Big((X - \mathbb{E}(X))^2\Big) = \sum_j p_j (x_j - \mathbb{E}(X))^2,$$

The intuition behind the variance is as follows:

- $x_j - \mathbb{E}(X)$ measures how far a particular x value is from the expected value

Intuition behind Variance

$$\mathrm{var}\,(X) \equiv \mathbb{E}\Big((X - \mathbb{E}(X))^2\Big) = \sum_j p_j (x_j - \mathbb{E}(X))^2,$$

The intuition behind the variance is as follows:

- $x_j - \mathbb{E}(X)$ measures how far a particular x value is from the expected value
- If we just used the difference, positives and negatives would cancel out, making it a bad measure of variability
- Therefore, we square the term to make it positive everywhere, which gives us a good measure of variability

Standard Deviation

The standard deviation is the square-root of the variance:

$$\operatorname{sd}(X) \equiv \sqrt{\operatorname{var}\left(X\right)}$$

Properties of Variance

In your own time, verify the following properties of variance:

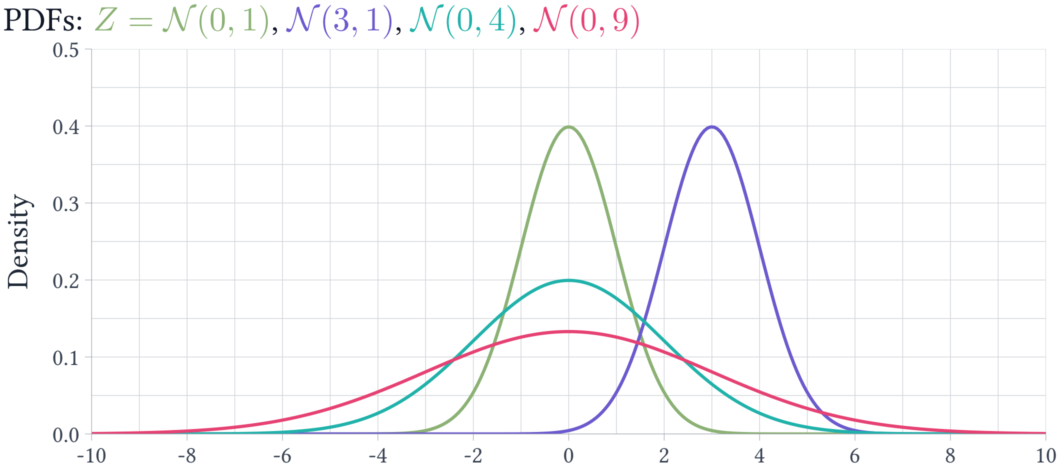
1. For any constant c, $\operatorname{var}(c) = 0$
2. For any constants a and b, $\operatorname{var}(aX + b) = a^2 \operatorname{var}(X)$

Normal Distribution

The normal distribution is one of the most important distributions in statistics. When a variable is normally distributed, we write $X \sim \mathcal{N}(\mu, \sigma^2)$, where $\mu = \mathbb{E}(X)$ and $\sigma^2 = \operatorname{var}(X)$.

- The normal distribution is symmetric and the density function of the normal distribution looks like a ‘bell’
- The parameter μ changes where the center of the distribution is and σ^2 determines how wide the distribution is

Example pdf of different normal distributions



Normal Distribution PDF

We can calculate probabilities of taking values $X \in [\underline{x}, \bar{x}]$ by integrating the area under the PDF. The PDF is given by

Normal Distribution PDF

We can calculate probabilities of taking values $X \in [\underline{x}, \bar{x}]$ by integrating the area under the PDF. The PDF is given by

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right).$$

Would you want to integrate this function? My guess is probably not.

Z-table and Standard Normal Distribution

Instead, we refer to the Z-table which calculates probabilities for the **standard normal distribution** denoted $Z = \mathcal{N}(0, 1)$.

The Z-table calculates $P(Z \leq z)$ for values of z ranging from -3 to 3.

Z-table and Standard Normal Distribution

Instead, we refer to the Z-table which calculates probabilities for the **standard normal distribution** denoted $Z = \mathcal{N}(0, 1)$.
The Z-table calculates $P(Z \leq z)$ for values of z ranging from -3 to 3.

Since we do not have a *standard* normal distribution, we can **standardize** our variable X to make it standard normal:

$$\frac{X - \mu}{\sigma} \sim Z = \mathcal{N}(0, 1)$$

Example: Standardizing and Using Z-table

Say we have $X \sim \mathcal{N}(10, 4)$ and we want to know $\mathbb{P}(X \leq 12)$. Then we can standardize our problem:

$$\mathbb{P}(X \leq 12) = \mathbb{P}\left(\frac{X - 10}{2} \leq \frac{12 - 10}{2}\right) = \mathbb{P}(Z \leq 1)$$

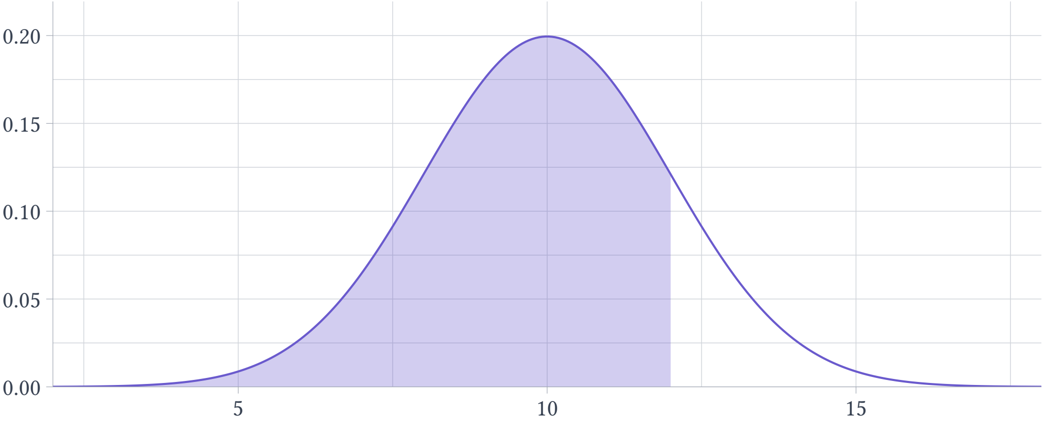
Example: Standardizing and Using Z-table

Say we have $X \sim \mathcal{N}(10, 4)$ and we want to know $\mathbb{P}(X \leq 12)$. Then we can standardize our problem:

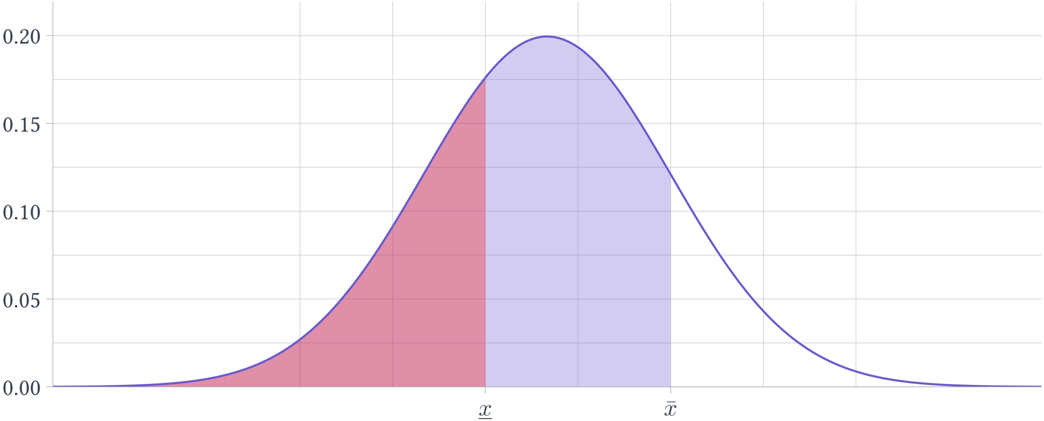
$$\mathbb{P}(X \leq 12) = \mathbb{P}\left(\frac{X - 10}{2} \leq \frac{12 - 10}{2}\right) = \mathbb{P}(Z \leq 1)$$

The last probability we can find in our Z-table by looking up the z-score of 1.

$\mathbb{P}[X \leq 12]$



$$\mathbb{P}[\underline{x} \leq X \leq \bar{x}] = \mathbb{P}[X \leq \bar{x}] - \mathbb{P}[\underline{x} \leq X]$$



Arbitrary intervals like $[\underline{x}, \bar{x}]$ can be calculated as $\mathbb{P}(X \leq \bar{x}) - \mathbb{P}(X \leq \underline{x})$.

Continuous Random Variables

Now we turn to continuous random variables, where random variables can take values on (parts of) the real number line (e.g. height).

- All our definitions and intuition still apply.
- In continuous land, we move from sums $\sum$ to integrals $\int$.

Probability Density Function

For continuous random variables, The PDF no longer represents the probability that X obtains the value x, since the probability of obtaining any particular value is 0

- Instead, the pdf gives a ‘relative likelihood’ that you take a value “near x”

Probability Density Function

For continuous random variables, The PDF no longer represents the probability that X obtains the value x, since the probability of obtaining any particular value is 0

- Instead, the pdf gives a ‘relative likelihood’ that you take a value “near x”

The pdf can be used to find the probability of being in a range of values $[\underline{x}, \bar{x}]$:

$$\mathbb{P}(X \leq x) = \int_{\underline{x}}^{\bar{x}} f_X(x) dx,$$

Cumulative Distribution Function

The PDF can be used to find the cumulative distribution function (CDF):

$$\mathbb{P}(X \leq x) = \int_{-\infty}^{x} f_X(x)dx,$$

Expectations for Continuous Random Variables

For continuous random variables, the expectation takes the form:

$$\mathbb{E}(X) = \int x * f_X(x) dx.$$

- This is similar to the discrete case, but we use an integral to ‘average’ the value times the density

Expectations for Continuous Random Variables

For continuous random variables, the expectation takes the form:

$$\mathbb{E}(X) = \int x * f_X(x) dx.$$

- This is similar to the discrete case, but we use an integral to ‘average’ the value times the density

All the properties of expectations listed above hold for continuous variables too (or mixtures of both).

Variance for Continuous Random Variables

Last, the variance is given by

$$\mathrm{var}\,(X) = \int (x - \mathbb{E}(X))^2 f_X(x) dx$$

Roadmap

Algebra Review

Linear Equations

Review of Statistics and Probability

Multiple Random Variables

Statistical Inference

Joint Distributions

In this class, we care about how variables *relate to one another*. For example:

- Do sales grow and shrink with a consumer’s age?
- Is height an important predictor of basketball success?
- Is the sale associated with a large increase in sales?

Joint Distributions

In this class, we care about how variables *relate to one another*. For example:

- Do sales grow and shrink with a consumer's age?
- Is height an important predictor of basketball success?
- Is the sale associated with a large increase in sales?

To answer these questions, we need to know about the joint distribution between two variables X and Y.

Joint Probability Density Function

The joint probability density function is denoted by $f_{X,Y}(x, y)$:

In the discrete case, it is the probability that $X = x$ *and* $Y = y$ *in the same trial*:

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

Conditional Probability Density Function

We can also think about conditioning on the value of one of the random variables

Take Y to be *fixed* to some value y. Then we could ask about the distribution of X *within trials where* $Y = y$:

$$f_{X|Y}(x|y) = \mathbb{P}(X = x \mid Y = y)$$

Conditional Probability Density Function

We can also think about conditioning on the value of one of the random variables

Take Y to be *fixed* to some value y. Then we could ask about the distribution of X *within trials where* $Y = y$:

$$f_{X|Y}(x|y) = \mathbb{P}(X = x \mid Y = y)$$

- Note we use the $|$ symbol to note 'conditioning' on some random variable's realization

Conditional Probability Density Function

We can also think about conditioning on the value of one of the random variables

Take Y to be *fixed* to some value y. Then we could ask about the distribution of X *within trials where* $Y = y$:

$$f_{X|Y}(x|y) = \mathbb{P}(X = x \mid Y = y)$$

- Note we use the | symbol to note ‘conditioning’ on some random variable’s realization
- In our example, we learn that $Y = y$ for that trial and then ask about the (conditional) probability that $X = x$

Bayes Rule

The Bayes Rule helps us translate between conditional pdfs and joint pdfs:

$$f_{X|Y}(x|y) = f_{X,Y}(x,y)/f_Y(y).$$

Bayes Rule

The Bayes Rule helps us translate between conditional pdfs and joint pdfs:

$$f_{X|Y}(x|y) = f_{X,Y}(x,y)/f_Y(y).$$

So say we have two discrete random variables and we want to know the probability $X = x$ conditional on $Y = y$. We can calculate it if we know:

1. the probability that $Y = y$
2. the joint probability that $Y = y$ *and* $X = x$

Covariance and Correlation

Similar to how we used statistics to summarize a single random variable, it is common to want to summarize how two variables are related to one another.

- For this, we use the covariance or the correlation between two variables (they are very similar)

Covariance

The covariance looks like the variance of a random variable:

$$\text{cov}(X, Y) = \mathbb{E}\left[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))\right]$$

The covariance intuitively measures whether X and Y move together:

- The covariance is positive if when X is above its mean, Y also tends to be above its mean; and when X is below its mean, Y also tends to be below its mean
- They “co-move” together

Covariance

The covariance is negative if when X is above its mean, Y tends to be below its mean and vice versa

- They move in opposite directions, but are still related!

Covariance

The covariance is negative if when X is above its mean, Y tends to be below its mean and vice versa

- They move in opposite directions, but are still related!
- In other words, if I know X was above its mean, then I would predict that Y is below its mean (knowing one gives me information on the other)

Alternative Formula for Covariance

With some algebra and the rules of expectations, we can see:

$$\begin{aligned}
\operatorname{cov}(X, Y) &= \mathbb{E}\left[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))\right] \\
&= \mathbb{E}\left[XY - \mathbb{E}(X)Y - X\,\mathbb{E}(Y) + \mathbb{E}(X)\,\mathbb{E}(Y)\right] \\
&= \mathbb{E}(XY) - \mathbb{E}(X)\,\mathbb{E}(Y) - \mathbb{E}(X)\,\mathbb{E}(Y) + \mathbb{E}(X)\,\mathbb{E}(Y) \\
&= \mathbb{E}(XY) - \mathbb{E}(X)\,\mathbb{E}(Y).
\end{aligned}$$

Correlation

The correlation is just a rescaled version of the covariance:

$$\operatorname{corr}(X, Y) = \frac{\operatorname{cov}(X, Y)}{\operatorname{sd}(X)\,\operatorname{sd}(Y)}.$$

The correlation is designed to always be between -1 and 1 (because of the rescaling):

- It is more popular since people are used to thinking about correlations.
- A correlation close to 1 and/or -1 is a *very strong* relationship.

Limitations of Covariance and Correlation

It is important to know that covariance and correlation only measure a *linear* relationship between X and Y:

- If the function connecting X and Y is non-linear, then the correlation is a bad summary statistic of the relationship between them.
- This is similar to how the mean is a bad measure for highly skewed data.

Independence

Two random variables are said to be independent if knowing information about the realization of one of them tells you nothing about the realization of the other one.

Independence

Two random variables are said to be independent if knowing information about the realization of one of them tells you nothing about the realization of the other one.

For example, if I told you the day was Sunday ($X =$ Sunday), then you would not have any better prediction about whether it is raining ($Y = 1$)

- The two random variables are independent

Independence and Conditional Probability

When X and Y are independent, this can be summarized by:

$$f_{X|Y}(x|y) = f_X(x).$$

Independence and Conditional Probability

When X and Y are independent, this can be summarized by:

$$f_{X|Y}(x|y) = f_X(x).$$

This also implies that when X and Y are independent,

$$F_{X,Y}(x,y) = F_X(x)F_Y(y).$$

Properties of Independent Random Variables

Using the definition of expectations, we can derive the following when X and Y are independent:

- $\mathbb{E}(XY) = \mathbb{E}(X)\,\mathbb{E}(Y)$
- $\operatorname{cov}(X, Y) = 0$

Properties of Independent Random Variables

Using the definition of expectations, we can derive the following when X and Y are independent:

- $\mathbb{E}(XY) = \mathbb{E}(X)\,\mathbb{E}(Y)$
- $\operatorname{cov}(X, Y) = 0$
- $\operatorname{var}(aX + bY) = a^2 \operatorname{var}(X) + b^2 \operatorname{var}(Y)$

The last fact comes from the fact that for *all* random variables $\operatorname{var}(aX + bY) = a^2 \operatorname{var}(X) + b^2 \operatorname{var}(Y) + 2ab\operatorname{cov}(X, Y)$, but the last term is zero from independence.

Roadmap

Algebra Review

Linear Equations

Review of Statistics and Probability

Multiple Random Variables

Statistical Inference

Statistical Inference

Statistical Inference is the procedure of using a random sample of observations from a population to try and learn some summary of the population distribution of the data.

Statistical Inference

Statistical Inference is the procedure of using a random sample of observations from a population to try and learn some summary of the population distribution of the data.

To be more specific, there is some statistic of the population distribution (e.g. it's expectation, it's variance, the 80th percentile, etc.)

- We will call the target statistic θ

Estimating Population Statistics

We do not observe the full population; only a random sample, $X_1, \ldots, X_n$. We take our data and calculate an estimate of the population statistic θ

- The calculation we choose is called our estimator.

Estimating Population Statistics

We do not observe the full population; only a random sample, $X_1, \ldots, X_n$. We take our data and calculate an estimate of the population statistic θ

- The calculation we choose is called our estimator.

For example, we can use the sample average from a set of observations to infer about the expectation of the population. To do so, we chose the sample average as our estimator:

$$\underbrace{\bar{X}}_{\text{Estimator}} \equiv \frac{1}{n}\sum_{i=1}^{n} X_i$$

Sample Distribution of an Estimator: Thought Experiment

To understand the sample distribution of an estimator, we need a thought experiment called repeated sampling.

Sample Distribution of an Estimator: Thought Experiment

To understand the sample distribution of an estimator, we need a thought experiment called repeated sampling.

In reality, we only have a single sample of size n. But, the thought experiment is is to imagine collecting multiple samples and calculating the estimator for each sample:

- The distribution of estimates across repeated sampling is called the **sample distribution**

Sample Distribution of an Estimator: Thought Experiment

To understand the sample distribution of an estimator, we need a thought experiment called repeated sampling.

In reality, we only have a single sample of size n. But, the thought experiment is is to imagine collecting multiple samples and calculating the estimator for each sample:

- The distribution of estimates across repeated sampling is called the **sample distribution**

The sample distribution can be used to quantify how much our estimator varies across different samples of the same size, n

Estimators

For any statistic, there are many different estimators. Therefore, we have various ways to discuss what makes a good estimator.

Property 1: Unbiasedness

An estimator W is unbiased for θ if:

$$\mathbb{E}(W) = \theta.$$

This means that, on average, the estimator equals the population statistic when using repeated sampling.

- It does not imply that the estimate always equals the population statistic.

Property 1: Unbiasedness

An estimator W is unbiased for θ if:

$$\mathbb{E}(W) = \theta.$$

This means that, on average, the estimator equals the population statistic when using repeated sampling.

- It does not imply that the estimate always equals the population statistic.

For instance, the sample mean is an unbiased estimate of the population mean.

Property 2: Consistency

Let W_n be an estimator of θ based on a sample size n. The estimator is consistent if, as n increases, $W_n \to \theta$ and the variance of W_n approaches 0.

Property 2: Consistency

Let W_n be an estimator of θ based on a sample size n. The estimator is consistent if, as n increases, $W_n \to \theta$ and the variance of W_n approaches 0.

- This means the estimator becomes more precise and centers around the population statistic.

Property 3: Efficiency

If W_1 and W_2 are two unbiased estimators, we say W_1 is more efficient than W_2 if:

$$\operatorname{var}(W_1) < \operatorname{var}(W_2)$$

This indicates that W_1 has a smaller variance, making it a more precise estimator

Describing Uncertainty of an Estimator

Since the estimator does not always equal its population statistic in repeated sampling, we need to describe the uncertainty of an estimator.

- To do this, we require an estimate of the variability of our estimator in repeated sampling

Normal Distribution and Confidence Intervals

It turns out that the *sample distribution* of many estimators is approximately normally distributed

- Makes it easier to summarize uncertainty around the estimator

Normal Distribution and Confidence Intervals

It turns out that the *sample distribution* of many estimators is approximately normally distributed

- Makes it easier to summarize uncertainty around the estimator

A confidence interval is constructed to describe the level of certainty

- Centered on the estimate, with a width that describes a range of values the population statistic could be

Sample Mean Distribution

Let $\{Y_1, \ldots, Y_n\}$ be a random sample. The sample mean is approximately distributed $\mathcal{N}(\mu, \sigma^2/n)$ where $\mu = \mathbb{E}(Y)$ and $\sigma^2 = \operatorname{var}(Y)$.

Sample Mean Distribution

Let $\{Y_1, \ldots, Y_n\}$ be a random sample. The sample mean is approximately distributed $\mathcal{N}(\mu, \sigma^2/n)$ where $\mu = \mathbb{E}(Y)$ and $\sigma^2 = \operatorname{var}(Y)$.

We can construct a 95% confidence interval as:

$$\mathbb{P}\left(-1.96 < \frac{\bar{Y} - \mu}{\sigma/\sqrt(n)} < 1.96\right) = 0.95$$

$$\implies \mathbb{P}\left(\bar{Y} - 1.96 * \frac{\sigma}{\sqrt(n)} < \mu < \bar{Y} + 1.96\frac{\sigma}{\sqrt(n)}\right) = 0.95$$

Confidence Interval

That is, in repeated sampling, there is a 95% probability (95 out of 100 samples of size n) that μ falls within our confidence interval:

$$\left[\bar{Y} - 1.96 * \frac{\sigma}{\sqrt(n)}, \bar{Y} + 1.96\frac{\sigma}{\sqrt(n)}\right]$$

Confidence Interval

That is, in repeated sampling, there is a 95% probability (95 out of 100 samples of size n) that μ falls within our confidence interval:

$$\left[\bar{Y} - 1.96 * \frac{\sigma}{\sqrt(n)}, \bar{Y} + 1.96\frac{\sigma}{\sqrt(n)}\right]$$

Since we do not observe σ, we will estimate it using the following estimator, called the sample standard deviation of y:

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

Critical Values and Confidence Level

The values -1.96 and 1.96 are the critical values for the 2.5th percentile and 97.5th percentile of the standard normal distribution, corresponding to a 95% confidence level.

Critical Values and Confidence Level

The values -1.96 and 1.96 are the critical values for the 2.5th percentile and 97.5th percentile of the standard normal distribution, corresponding to a 95% confidence level.

To increase the confidence level, we need to use larger critical values. For example, for a 99% confidence level, we use the 0.5th percentile and 99.5th percentile of the normal distribution for our critical values.

Z-score and Standard Normal Distribution

This procedure works because $\bar{Y} \sim \mathcal{N}(\mu, \sigma^2/n)$ and therefore our Z**-score**, $\frac{\bar{Y}-\mu}{\sigma/\sqrt{(n)}}$, is approximately distributed as the standard normal distribution.

- Therefore using the standard normal distribution to chose critical values works

Hypothesis Testing

In addition to confidence intervals, it is common to perform hypothesis tests.

Simply put, we want to test whether evidence is consistent with a null hypothesis that θ equals some value (e.g. we could hypothesize that the population mean height is 5 ft. 8 in.).

- We then would calculate our sample mean and if it is “too far” from the null hypothesis population mean, then we say the evidence rejects the null hypothesis.

Determining "Too Far"

How far is “too far” away from the null hypothesis? That depends on

1. how noisy our estimate is (the sample distribution), and
2. how confident we want to be in rejecting the null

Test Statistics

We will use our sample's estimate to calculate a test statistic. The test statistic (a random variable) will be referred to as T and a realized value as t.

When the null hypothesis is true, the test statistic will be distributed as $\mathcal{N}(0, v)$ (normally distributed with mean 0 and some variance v).

Test Statistics

We will use our sample's estimate to calculate a test statistic. The test statistic (a random variable) will be referred to as T and a realized value as t.

When the null hypothesis is true, the test statistic will be distributed as $\mathcal{N}(0, v)$ (normally distributed with mean 0 and some variance v).

- Therefore, we expect T to be very close to zero (if the null is true)
- So, if we find a value of $T = t$ that is "very far" away from zero, then we find evidence against the null hypothesis

p-Value

Given a level of confidence (e.g. 95%) and the variance v, we can use properties of the normal distribution to determine the probability that a draw from the $T \sim \mathcal{N}(0, v)$ distribution would be $\geq \left|\hat{t}\right|$

This is called the p**-value**

- In words, the p-value tells us, "assuming the null hypothesis is true, how often would we expect to see a value as large or larger than the one we *did observe in our sample*"
- If that probability is small, then we will reject the null.

Significance Level

In particular, given our level of confidence (e.g. 95%), define the **significance level** as $\alpha = 1 -$ the level of confidence (e.g. 5%)

We reject the null if the p-value, p is smaller than the significance level

- Typically, we will use $\alpha = 0.05$

Example: Hypothesis Testing for Sample Mean

Let $\{Y_1, \dots, Y_n\}$ be a random sample. The sample mean is approximately distributed $\mathcal{N}(\mu, \sigma^2/n)$ where $\mu = \mathbb{E}(Y)$ and $\sigma^2 = \operatorname{var}(Y)$

We will test the null hypothesis $H_0 : \mu = \mu_0$ for some proposed value μ_0

Example: Hypothesis Testing for Sample Mean

Then, our test statistic is given by

$$T = \frac{\bar{X} - \mu_0}{\hat{\sigma}/\sqrt(n)},$$

where $\hat{\sigma}$ is the square root of our estimate of the variance of Y

- Assuming the null hypothesis is true, H_0, then $T \sim \mathcal{N}(0, 1)$.

Example: Hypothesis Testing for Sample Mean

Therefore, the p-value is given by

$$p = \mathbb{P}(T \geq |t|) = 2 * \mathbb{P}(Z \leq -|t|),$$

where Z is the standard normal random variable.

Example: Hypothesis Testing for Sample Mean

Therefore, the p-value is given by

$$p = \mathbb{P}(T \geq |t|) = 2 * \mathbb{P}(Z \leq -|t|),$$

where Z is the standard normal random variable.

- You can find this probability by looking up $-|t|$ in the Z-table.

Rejection Region

Last, we will discuss the **rejection region** for a given *null hypothesis*:

This is defined as the set of values of our estimate where we would reject the null hypothesis for a given significance level, α.

Example: Hypothesis Testing for Sample Mean

The rejection region can be found, most simply, by forming a confidence interval around the null hypothesis. Any value *outside this interval* will be rejected.

Example: Hypothesis Testing for Sample Mean

The rejection region can be found, most simply, by forming a confidence interval around the null hypothesis. Any value *outside this interval* will be rejected.

Our confidence interval for $\alpha = 0.05$ is given by

$$\left[\mu_0 - 1.96 * \sigma/\sqrt{n}, \mu_0 + 1.96 * \sigma/\sqrt{n}\right]$$

Therefore our rejection region is $\bar{X} \leq \mu_0 - 1.96 * \sigma/\sqrt{n}$ or $\bar{X} \geq \mu_0 + 1.96 * \sigma/\sqrt{n}$.