Statistical Foundations of Data Science

Bayes Theorem and The Normal Distribution

University of the Witwatersrand

2025

Review Question

- Compute $\binom{10}{4}$

Lesson Plan

- Review Question

Lesson Plan

- Review Question
- Owed - Conditional distributions - Continuous

Lesson Plan

- Review Question
- Owed - Conditional distributions - Continuous
- The normal distribution

Lesson Plan

- Review Question
- Owed - Conditional distributions - Continuous
- The normal distribution
- The Central Limit Theorem

Lesson Plan

- Review Question
- Owed - Conditional distributions - Continuous
- The normal distribution
- The Central Limit Theorem
- The multivariate normal distribution

Review Question

- Compute $\binom{10}{4}$

Review Question

- Compute $\binom{10}{4}$
- Compute $\frac{10\times9\times8\times7}{4\times3\times2\times1}$

Review Question

- Compute $\binom{10}{4}$
- Compute $\frac{10\times9\times8\times7}{4\times3\times2\times1}$
- Compute $\frac{10\times9\times8\times7}{4\times3\times2}$

Review Question

- Compute $\binom{10}{4}$
- Compute $\frac{10\times9\times8\times7}{4\times3\times2\times1}$
- Compute $\frac{10\times9\times8\times7}{4\times3\times2}$
- Compute $\frac{10\times9\times7}{3}$

Review Question

- Compute $\binom{10}{4}$
- Compute $\frac{10\times9\times8\times7}{4\times3\times2\times1}$
- Compute $\frac{10\times9\times8\times7}{4\times3\times2}$
- Compute $\frac{10\times9\times7}{3}$
- Compute $10 \times 3 \times 7$

Review Question

- Compute $\binom{10}{4}$
- Compute $\frac{10\times9\times8\times7}{4\times3\times2\times1}$
- Compute $\frac{10\times9\times8\times7}{4\times3\times2}$
- Compute $\frac{10\times9\times7}{3}$
- Compute $10 \times 3 \times 7$
- 210

Normal Distribution

- Bell shaped distribution

Normal Distribution

- Bell shaped distribution
- $f(x) = \frac{1}{2\sqrt{\pi\sigma}} e^{\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$

Normal Distribution

- Bell shaped distribution
- $f(x) = \frac{1}{2\sqrt{\pi\sigma}} e^{\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$
- μ represents mean

Normal Distribution

- Bell shaped distribution
- $f(x) = \frac{1}{2\sqrt{\pi\sigma}} e^{\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$
- μ represents mean
- σ represents standard deviation.

Normal Distribution

- Bell shaped distribution
- $f(x) = \frac{1}{2\sqrt{\pi\sigma}} e^{\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$
- μ represents mean
- σ represents standard deviation.
- This is (provably) impossible to integrate analytically. So we use tables.

Normal Distribution

- Bell shaped distribution
- $f(x) = \frac{1}{2\sqrt{\pi\sigma}} e^{\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$
- μ represents mean
- σ represents standard deviation.
- This is (provably) impossible to integrate analytically. So we use tables.
- Depends on two parameters μ and σ so we make tables for the standard normal $\mu = 0$ and $\sigma = 1$

Normal Distribution

- Bell shaped distribution
- $f(x) = \frac{1}{2\sqrt{\pi\sigma}} e^{\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$
- μ represents mean
- σ represents standard deviation.
- This is (provably) impossible to integrate analytically. So we use tables.
- Depends on two parameters μ and σ so we make tables for the standard normal $\mu = 0$ and $\sigma = 1$
- We compute Z scores. Which are essentially just the number of standard deviations above the mean.

Normal Distribution

- Bell shaped distribution
- $f(x) = \frac{1}{2\sqrt{\pi\sigma}} e^{\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$
- μ represents mean
- σ represents standard deviation.
- This is (provably) impossible to integrate analytically. So we use tables.
- Depends on two parameters μ and σ so we make tables for the standard normal $\mu = 0$ and $\sigma = 1$
- We compute Z scores. Which are essentially just the number of standard deviations above the mean.
- Many things are actually normal. Because of the CLT.

Normal distribution - Example 1

- Goliath is 2.2 m tall! If the average height is 1.75m and the standard deviation is 15cm what proportion of the population is taller than Goliath?

Normal distribution - Example 1

- Goliath is 2.2 m tall! If the average height is 1.75m and the standard deviation is $15cm$ what proportion of the population is taller than Goliath?
- Goliath's Z-score is $\frac{X-\mu}{\sigma} = \frac{220-175}{15} = 3$. We use a table to see 0.9987 or 99.87 percent of the population is shorter.

Normal distribution - Example 2

- Albert E is disappointed to learn he got only 75 percent on his most recent test! However his teacher informs him that he was actually 3 standard deviations above the class average of 50 which cheers him up. Find the standard deviation.

Normal distribution - Example 2

- Albert E is disappointed to learn he got only 75 percent on his most recent test! However his teacher informs him that he was actually 3 standard deviations above the class average of 50 which cheers him up. Find the standard deviation.
- $$\begin{aligned} Z &= \frac{X-\mu}{\sigma} \\ 3 &= \frac{75-50}{\sigma} \\ \sigma &= 8.333 \end{aligned}$$

Normal distribution - Example 2

- For the next test Albert E is excited to learn that he was 4 standard deviations above the mean with a score of 83! He's surprised to learn however that the standard deviation was only 2 points. Find the mean score

Normal distribution - Example 2

- For the next test Albert E is excited to learn that he was 4 standard deviations above the mean with a score of 83! He's surprised to learn however that the standard deviation was only 2 points. Find the mean score
-

$$Z = \frac{X - \mu}{\sigma}$$
$$4 = \frac{83 - \mu}{2}$$
$$\mu = 75$$

Central Limit Theorem

- Roughly says that if we take n iid random variables then the sum (or mean) of these is (nearly) normal

Central Limit Theorem

- Roughly says that if we take n iid random variables then the sum (or mean) of these is (nearly) normal
- Well if we take n we'll get a fixed number but if we take n lots of times then the different sums are (nearly) normal.

Central Limit Theorem

- Roughly says that if we take n iid random variables then the sum (or mean) of these is (nearly) normal
- Well if we take n we'll get a fixed number but if we take n lots of times then the different sums are (nearly) normal.
- Nearly normal in the sense that the distribution of the means approximates a normal better and better as n gets larger.

Central Limit Theorem

- Roughly says that if we take n iid random variables then the sum (or mean) of these is (nearly) normal
- Well if we take n we'll get a fixed number but if we take n lots of times then the different sums are (nearly) normal.
- Nearly normal in the sense that the distribution of the means approximates a normal better and better as n gets larger.
- This also assumes that the distribution we're sampling from has a mean and variance. Most distributions do.

Central Limit Theorem

- More particularly if X_i have mean μ and variance σ^2 $\overline{X}$ is distributed $N(\mu, \frac{\sigma^2}{n})$. Same mean but standard deviation is divided by $\sqrt{n}$. For large n the sample mean is therefore pretty accurate.

Central Limit Theorem

- More particularly if X_i have mean μ and variance σ^2 $\overline{X}$ is distributed $N(\mu, \frac{\sigma^2}{n})$. Same mean but standard deviation is divided by $\sqrt{n}$. For large n the sample mean is therefore pretty accurate.
- There are variants of the CLT with relaxed assumptions! These include weak dependence within the samples.

Formal Statement of the Central Limit Theorem

Let $X_1, X_2, \ldots, X_n$ be a sequence of i.i.d. random variables with:

- Mean $\mathbb{E}[X_i] = \mu$,
- Variance $\mathrm{Var}(X_i) = \sigma^2$, assuming $0 < \sigma^2 < \infty$.

Define the standardized sum:

$$Z_n = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}}$$

Then as $n \to \infty$,

$$Z_n \to \mathcal{N}(0, 1).$$

Proof Outline

- We use characteristic functions: $\varphi_X(t) = \mathbb{E}[e^{itX}]$.
- The characteristic function is given by:

$$\varphi_X(t) = \mathbb{E}[e^{itX}].$$

- The characteristic function can be thought of as a clothesline to hang moments from. To seethis, we expand e^{itX} into its Taylor series:

$$e^{itX} = \sum_{k=0}^{\infty} \frac{(itX)^k}{k!}.$$

- Taking expectation term by term:

$$\varphi_X(t) = \sum_{k=0}^{\infty} \frac{(it)^k}{k!} \mathbb{E}[X^k].$$

Proof Outline

- Comparing with the power series of e^{itX}, we see:

$$\mathbb{E}[X^k] = \frac{d^k}{dt^k}\varphi_X(t)\bigg|_{t=0}.$$

- Thus, the characteristic function generates moments through differentiation.
- The characteristic function of a sum is given by:

$$\varphi_{S_n}(t) = \mathbb{E}[e^{itS_n}].$$

- To see this notice $S_n = X_1 + X_2 + \cdots + X_n$, we expand:

$$\varphi_{S_n}(t) = \mathbb{E}[e^{it(X_1+X_2+\cdots+X_n)}].$$

- Using the property of exponentials:

$$e^{it(X_1+X_2+\cdots+X_n)} = e^{itX_1}e^{itX_2}\dots e^{itX_n}.$$

- By independence, expectation distributes:

$$\mathbb{E}[e^{itX_1}e^{itX_2}\dots e^{itX_n}] = \mathbb{E}[e^{itX_1}]\mathbb{E}[e^{itX_2}]\dots\mathbb{E}[e^{itX_n}].$$

$\varphi_{S_n}(t) = (\varphi_X(t))^n$.
Expand $\varphi_X(t)$ around $t = 0$:

$$\varphi_X(t) = 1 + it\mu - \frac{t^2\sigma^2}{2} + o(t^2).$$

Proof outline

$$\varphi_{S_n}(t) = \left(1 + it\mu - \frac{t^2\sigma^2}{2} + o(t^2)\right)^n.$$

Using $(1 + x)^n \approx e^{nx}$ for small x:

$$\varphi_{S_n}(t) \approx e^{n(it\mu - \frac{t^2\sigma^2}{2})}.$$

Proof outline

Define $Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}}$. The characteristic function of Z_n is:

$$\varphi_{Z_n}(t) = e^{-\frac{t^2}{2}}.$$

This is exactly the characteristic function of $\mathcal{N}(0, 1)$.

Conclusion

- By Lévy's continuity theorem, $Z_n \xrightarrow{d} \mathcal{N}(0,1)$.
- This completes the proof of the Central Limit Theorem.

Multivariate Normal Distribution

- Multivariate meaning it's a distribution on more than one variable.

Multivariate Normal Distribution

- Multivariate meaning it's a distribution on more than one variable.
- Defined by a mean vector and covariance matrix

Multivariate Normal Distribution

- Multivariate meaning it's a distribution on more than one variable.
- Defined by a mean vector and covariance matrix
- If the covariance matrix is diagonal then it's a bunch of independent normals.

Multivariate Normal Distribution

- Multivariate meaning it's a distribution on more than one variable.
- Defined by a mean vector and covariance matrix
- If the covariance matrix is diagonal then it's a bunch of independent normals.
- Will dominate our discussion of multivariate regression

Multivariate Normal Distribution

- Multivariate meaning it's a distribution on more than one variable.
- Defined by a mean vector and covariance matrix
- If the covariance matrix is diagonal then it's a bunch of independent normals.
- Will dominate our discussion of multivariate regression
- Linear combinations of all variables are still normal

Multivariate Normal Distribution

- Multivariate meaning it's a distribution on more than one variable.
- Defined by a mean vector and covariance matrix
- If the covariance matrix is diagonal then it's a bunch of independent normals.
- Will dominate our discussion of multivariate regression
- Linear combinations of all variables are still normal
- All conditional distributions are normal.

Conditional Probability

- A person is drawn from the world's population. What do you think the probability is that they're over 1.8m in height?

Conditional Probability

- A person is drawn from the world's population. What do you think the probability is that they're over 1.8m in height?
- What if I say that they are a male aged between 25 and 30?

Conditional Probability

- A person is drawn from the world's population. What do you think the probability is that they're over 1.8m in height?
- What if I say that they are a male aged between 25 and 30?
- The point is that some knowledge of a dataset tells us something about the population as a whole.

Classic Example

- You take a covid-test. The test comes back correctly 95 percent of the time. What's the probability that you have it?

Classic Example

- You take a covid-test. The test comes back correctly 95 percent of the time. What's the probability that you have it?
- Naively we might think 0.95 but it turns out that it depends!

Classic Example

- You take a covid-test. The test comes back correctly 95 percent of the time. What's the probability that you have it?
- Naively we might think 0.95 but it turns out that it depends!
- Example. What if you're in 2017 and someone time travelled back with a test? Or you're on the International Space Station?

Classic Example

- You take a covid-test. The test comes back correctly 95 percent of the time. What's the probability that you have it?
- Naively we might think 0.95 but it turns out that it depends!
- Example. What if you're in 2017 and someone time travelled back with a test? Or you're on the International Space Station?
- It turns out that your prior probability matters!

Classic Example

- $$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B|A)$$
- $$\mathbb{P}(A)\mathbb{P}(B|A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)}$$

Example

- Well that's great but what if I actually want to calculate a probability? Let's say that the prevalence is 10 percent.

Example

- Well that's great but what if I actually want to calculate a probability? Let's say that the prevalence is 10 percent.
- Two kinds of people get back positive covid tests. Those who have covid and for whom the test works. These make up $0.1 \times 0.95 = 0.095$ of the population. The other kind are those who don't have covid but who the test failed for. These make up $0.9 \times 0.05 = 0.045$ of the population.

Example

- Well that's great but what if I actually want to calculate a probability? Let's say that the prevalence is 10 percent.
- Two kinds of people get back positive covid tests. Those who have covid and for whom the test works. These make up $0.1 \times 0.95 = 0.095$ of the population. The other kind are those who don't have covid but who the test failed for. These make up $0.9 \times 0.05 = 0.045$ of the population.
- So your actual probability of having the disease given the test is $\frac{0.095}{0.095+0.045} = 0.67857142857$.

Example

- Well that's great but what if I actually want to calculate a probability? Let's say that the prevalence is 10 percent.
- Two kinds of people get back positive covid tests. Those who have covid and for whom the test works. These make up $0.1 \times 0.95 = 0.095$ of the population. The other kind are those who don't have covid but who the test failed for. These make up $0.9 \times 0.05 = 0.045$ of the population.
- So your actual probability of having the disease given the test is $\frac{0.095}{0.095+0.045} = 0.67857142857$.
- These numbers can be tweaked a lot. Again the extreme example is testing for a non-existent disease.

Dice Example

- You roll a standard dice. What's the probability of rolling an even number?

Dice Example

- You roll a standard dice. What's the probability of rolling an even number?
- Of course it's $\frac{3}{6} = \frac{1}{2} = 0.5$

Dice Example

- You roll a standard dice. What's the probability of rolling an even number?
- Of course it's $\frac{3}{6} = \frac{1}{2} = 0.5$
- What's the probability of rolling an even number given you rolled a prime?

Dice Example

- You roll a standard dice. What's the probability of rolling an even number?
- Of course it's $\frac{3}{6} = \frac{1}{2} = 0.5$
- What's the probability of rolling an even number given you rolled a prime?
- Well the possible primes are 2, 3 and 5. So we know we rolled one of those. Hence $\frac{1}{3}$

Dice Example

- You roll a standard dice. What's the probability of rolling an even number?
- Of course it's $\frac{3}{6} = \frac{1}{2} = 0.5$
- What's the probability of rolling an even number given you rolled a prime?
- Well the possible primes are 2, 3 and 5. So we know we rolled one of those. Hence $\frac{1}{3}$
- Generally getting some information puts us in a subset of the sample space.

More Generally

- $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B|A)$

More Generally

- $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B|A)$
- So $\mathbb{P}(B|A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)}$

More Generally

- $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B|A)$
- So $\mathbb{P}(B|A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)}$
- The exception is when $\mathbb{P}(A) = 0$, this happens with continuous models. There are entirely natural and good and rigorous ways to handle this, for the most part we'll just notice it's an issue.

More Generally

- $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B|A)$
- So $\mathbb{P}(B|A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)}$
- The exception is when $\mathbb{P}(A) = 0$, this happens with continuous models. There are entirely natural and good and rigorous ways to handle this, for the most part we'll just notice it's an issue.
- When we get to the regression part of this course we'll do a lot of this. For example we might consider the conditional distribution of someone's weight given their height, while the probability of the height being exactly $1.81231238290438238223928 3m$ is zero.

Example for paradox

- Let X and Y be independent uniform continous random variables of $[0, 1]$

Example for paradox

- Let X and Y be independent uniform continous random variables of $[0, 1]$
- Find $\mathbb{P}(XY > 0.5|X = 0.8)$

Example for paradox

- Let X and Y be independent uniform continous random variables of $[0, 1]$
- Find $\mathbb{P}(XY > 0.5|X = 0.8)$
- $\mathbb{P}(0.8Y > 0.5) = \mathbb{P}(Y > 0.5/0.8) = 0.375$

Example for paradox

- Let X and Y be independent uniform continous random variables of $[0, 1]$
- Find $\mathbb{P}(XY > 0.5|X = 0.8)$
- $\mathbb{P}(0.8Y > 0.5) = \mathbb{P}(Y > 0.5/0.8) = 0.375$
- Here the probability that $X = 0.8$ is zero but you know what to do.