Module C: Understanding the Data-Generating Process

Slide Deck C9:

The Normal Distribution

The section in which we learn about another named continuous distribution. The Normal distribution is the most important continuous distribution in terms of usefulness... as we shall see with the Central Limit Theorem.

Today's Objectives

By the end of this slidedeck, you should

1. discuss the differences between the Uniform, Exponential, and Normal distributions;
2. express the difference between 'normal' and 'Normal'
3. understand the importance of the probability density function (pdf)
4. understand the importance of the cumulative distribution function (CDF)
5. be able to determine the
   - sample space,
   - mean, variance,
   - median, quantiles (percentiles)

Characteristics of the Gaussian distribution

Here are some important features of the Normal distribution:

- It is named after Carl Friedrich Gauss who advocated for its use (1809, 1794?)
- The Normal distribution arises from modeling observed randomness (errors) in astronomical and geodetic data
- It arises in variables that have a specific expected value, but demonstrate minor random variation
- It has two parameters: μ (expected value) and σ (st. dev.)
- It is the most important distribution in statistics because of the Central Limit Theorem

Characteristics of the Gaussian distribution

For the record, if

$$X \sim \mathcal{N}(\mu, \sigma)$$

then the probability density function (pdf) is

$$f(x;\ \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}}\ \mathbb{e}^{-\frac{1}{2}\ \frac{(x-\mu)^2}{\sigma^2}}$$

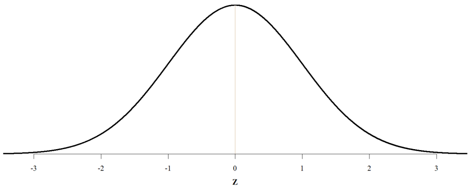
This pdf depends on values of μ and σ. Note, however, that we can standardize X (*à la* Slidedeck B3 and the **z-score**) so that the distribution no longer depends on either μ or σ.

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

It is *this* standard Normal distribution that textbooks tabulate.

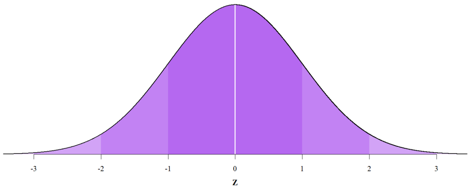
Characteristics of the Gaussian distribution

The following is a graph of the standard Normal pdf, $\phi(z)$,



Characteristics of the Gaussian distribution

The following is a graph of the standard Normal pdf, $\phi(z)$,



Calculating Features

Recall that the Normal distribution is a continuous distribution

- not rectangular
- the CDF is not calculable

Thus, calculating probabilities requires either a Table or a computer. By tradition, "Table I" tabulates the CDF of the standard Normal distribution, $\Phi(z)$. However, computer calculations are easier to perform. . . and are more accurate and precise. Use them.

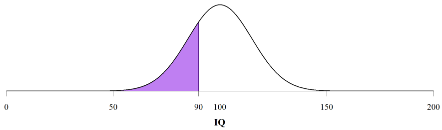
Here, let $X \sim \mathcal{N}(m, s)$ with $\mathbb{P}[\ X \leq x_p\ ] = p$:

Cumulative probability, p	`pnorm(x, m,s)`
Quantile, x_p	`qnorm(p, m,s)`

Example I: Intelligence Quotient

Example

By design, intelligence quotient (IQ) measures in the United States follow a Normal distribution with $\mu = 100$ and $\sigma = 15$. What proportion of the US population has an IQ less than 90?



Example I: Intelligence Quotient

Solution. We are asked to calculate p such that

$$\mathbb{P}[\ X < 90\ ] = p$$

given that $X \sim \mathcal{N}(100, 15)$.

This can be calculated using this `R` code:

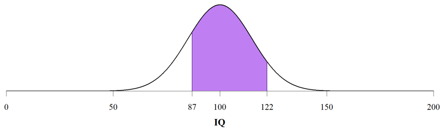
```
pnorm(90, m=100,s=15)
```

Thus, the proportion of people in the United States with IQ scores less than 90 is 0.2524925, which is about 25%.

Example II: Intelligence Quotient

Example

By design, intelligence quotient (IQ) measures in the United States follow a Normal distribution with $\mu = 100$ and $\sigma = 15$. What proportion of Americans have an IQ score between 87 and 122?



Example II: Intelligence Quotient

Solution. We are asked to calculate p such that

$$\mathbb{P}[\ 87 < X < 122\ ] = p$$

given that $X \sim \mathcal{N}(100, 15)$.

This can be calculated using this R code:

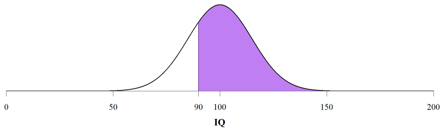
```
pnorm(122, m=100,s=15) - pnorm(87, m=100,s=15)
```

Thus, the proportion of Americans with an IQ score between 87 and 122 is 0.7357043*, approximately 75%.*

Example III: Intelligence Quotient

Example

By design, intelligence quotient (IQ) measures in the United States follow a Normal distribution with $\mu = 100$ and $\sigma = 15$. What proportion of Americans have an IQ score above 90?



Example III: Intelligence Quotient

Solution. We are asked to calculate p such that

$$\mathbb{P}[\ 90 > X\ ] = p$$

given that $X \sim \mathcal{N}(100, 15)$.

This can be calculated using this R code:

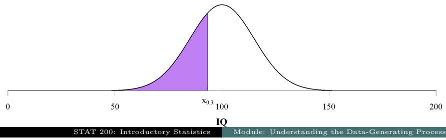
```
1 - pnorm(90, m=100,s=15)
```

Thus, the proportion of Americans with an IQ score above 90 is 0.7475075, *approximately 75%.*

Example IIII: Intelligence Quotient

Example

By design, intelligence quotient (IQ) measures in the United States follow a Normal distribution with $\mu = 100$ and $\sigma = 15$. What IQ value separates the lower 30% from the upper 70%?



Example IIII: Intelligence Quotient

Solution. We are asked to calculate x_p such that

$$\mathbb{P}[\ X < x_p\ ] = 0.30$$

given that $X \sim \mathcal{N}(100, 15)$.

This can be calculated using this R code:

```
qnorm(0.30, m=100,s=15)
```

Thus, the IQ score that separates the lower 30% from the upper 70% is the 30^{th} percentile: 92.13.

> *Approximately 30% of Americans have an IQ score less than* 92.13 *and approximately 70% of Americans have an IQ score of more than* 92.13.

Today's Summary

Now that we have concluded this lecture, you should be able to

1. discuss the differences between the Uniform, Exponential, and Normal distributions;
2. express the difference between 'normal' and 'Normal'
3. understand the importance of the probability density function (pdf)
4. understand the importance of the cumulative distribution function (CDF)
5. be able to determine the
    - sample space,
    - mean, variance,
    - median, quantiles (percentiles)

Useful R Functions

In this slide deck, we covered these R functions regarding the Normal distribution:

- `dnorm(x, m,s)` is the density function, $f(x)$
- `pnorm(x, m,s)` is the CDF $F(x) = \mathbb{P}[\ X \leq x\ ]$
- `qnorm(p, m,s)` is the quantile function $\mathbb{P}[\ X \leq x\ ] = p$
- `rnorm(n, m,s)` gives a random sample from this distribution

Please do not forget to use the `allProbabilities` document that lists all of the probability functions in R.

Supplemental Activities

The following are supplements for the topics covered today.

- SCA 6a is for continuous distributions like the Normal.

Note that you can access all Statistical Computing Activities here:
`https://www.kvasaheim.com/courses/stat200/sca/`

In addition to the SCA, **Laboratory Activity C** is helpful for learning how to handle some continuous distributions (including the Normal distribution). The lab actually illustrates the Central Limit Theorem.
`https://www.kvasaheim.com/courses/stat200/labs/`

Supplemental Readings

The following are some readings that may be of interest to you in terms of understanding continuous distributions, including the Normal:

- Hawkes Learning: Sections 6.1–6.4
- Intro to Modern Statistics: None
- R for Starters: Appendix B.3, C.1, and C.2
- Wikipedia: Gaussian Distribution
  Normal Distribution