Week 4 4.2 Note

Lecture 10: Normal Curve (Model)

1. Normal Curve

- Why is it important?
  - It approximates many *natural phenomenon*.
  - It can model data caused by *a large number of independent variables*.

> The **general** normal curve (X) has **any mean and SD**.

> The **standard** normal curve (Z) has **mean 0 and SD 1**.

- The normal curve formula is:
  - The formula for the General Normal Curve is

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad \text{for } x \in (-\infty, \infty)$$

where μ and σ are the (population) mean and SD respectively.

2. Area under Normal Curve

- Approximating Histogram by a Normal Curve
  - If the normal curve fits the histogram, we can use **the area under normal curve** as an approximation to **the area under the histogram**.
- How to calculate the area under a *standard normal curve*?
  - `pnorm()` works out the lower **tail area**.
    - ex. `pnorm(x, lower.tail=F)` works out the *right* tail area (which is `upper tail`).
  - For *intervals*, we do `pnorm()-pnorm()`.
    - ex. `pnorm(0.8)-pnorm(0.3)` = area between 0.3 and 0.8
- How to calculate the area under a *general normal curve*?
  - `pnorm(x,y,z)`, where x is the *scale*, y is the *mean*, z is the *standard deviation*.

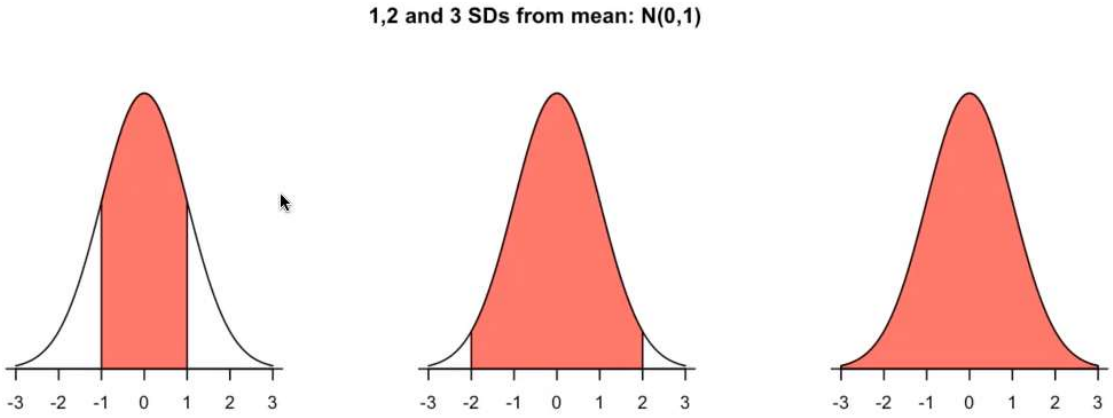
3. Special Properties of the Normal Curve

1. All normal curves satisfy the **68%-95%-99.7% rule**.

- Which is:

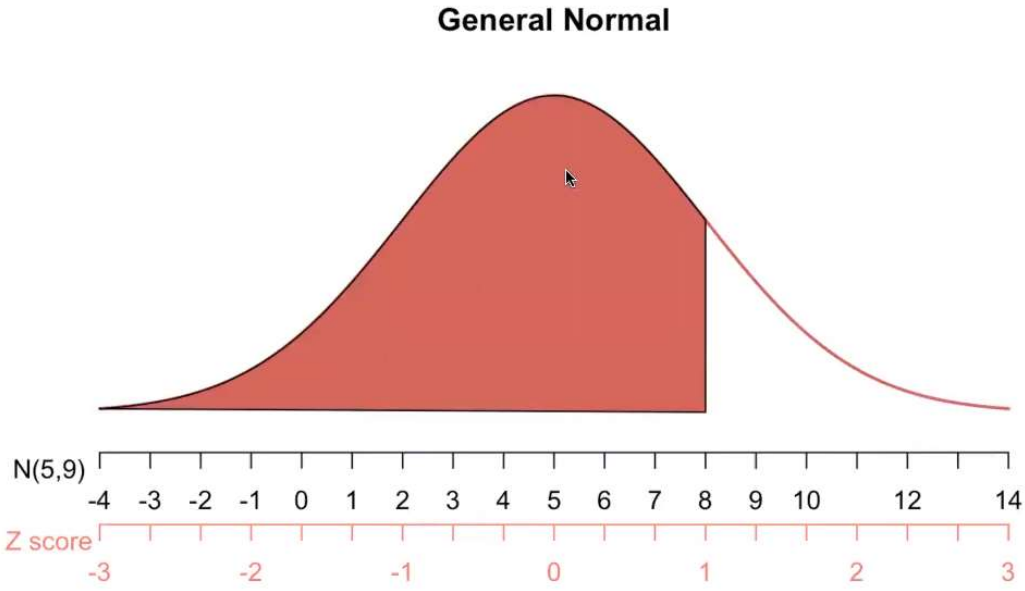
percentage of Data	Distance from Mean
68%	Within 1 SD
95%	Within 2 SDs
99.7%	Within 3 SDs

- Visually:

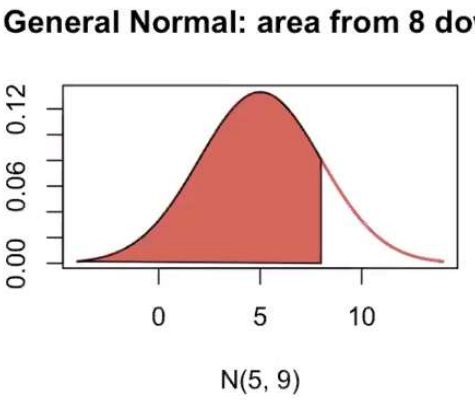


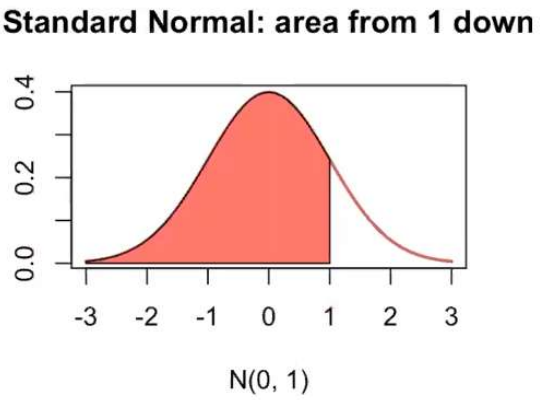
2. Any GN can be rescaled into SN.

- Steps:
    1. standardize the **thresholds** (data points) of the general using **standard units** (relevant points on the standard) = (data point - mean) / SD
    2. The new standard units are the thresholds of the new standard.
- ex.1

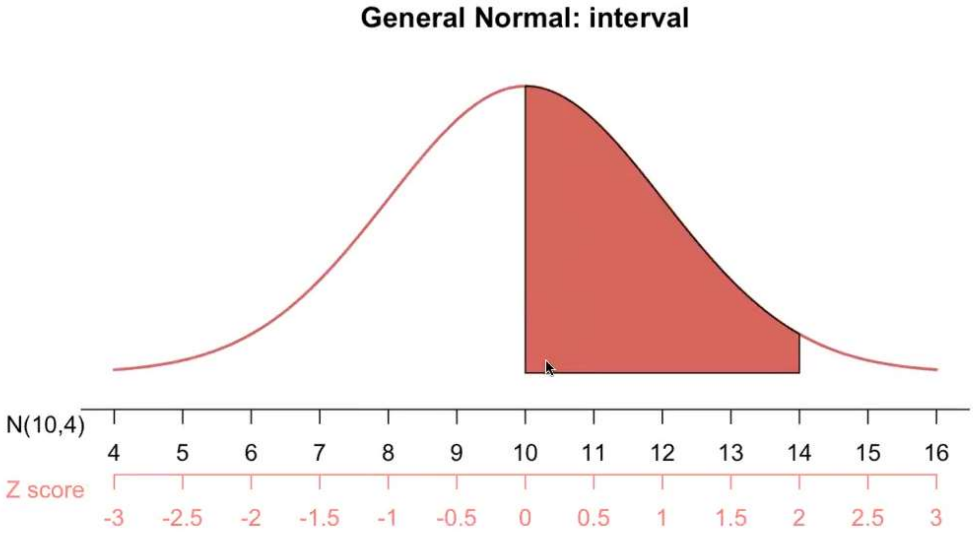


- Consider the point = 8.
- So the z score is $\frac{8-5}{3} = 1$.

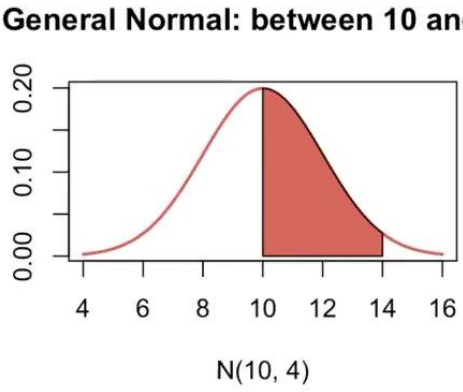


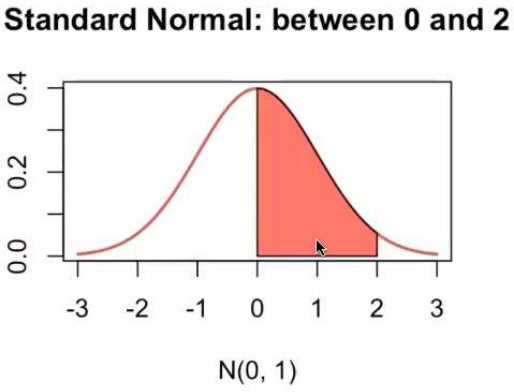


- ex.2



- Here the lower point is 10 and the upper point is 14.
- So the z scores are $z_1 = \frac{10-10}{2} = 0$ and $z_2 = \frac{14-10}{2} = 2$.





Lecture 11: Measurement Error

1. Measurement

- An individual measurement often differs from the exact value.
  - Individual Measurement = Exact Value + Chance Error + Bias

2. Chance Error

- Measurement will always turn out differently due to *chance error*.
- To estimate the chance error, we **replicate** the measurement under the same conditions and calculate the **standard deviation**.

3. Outliers

> Outliers: A small part of **extreme measurements** in large series of measurements.
>
> - Standard units (Distance from data from the mean) can tell the outliers.

4. Bias

> Bias (systematic error): **A constant amount** added to or subtracted from **each measurement**.

- Bias can be deliberate or accidental.

Lecture 12: Reproducible Reports

---

1. R Markdown

> R Script: A text file which saves the R code and comments "#".

> R Markdown: An authoring framework for data science which produces dynamic and interactive documents with R

- It combines:
  - Chunks of text
  - Embedded code
  - Latex
- Customize Code Chunk:
  - Echo = F: Don't show code (only show the output)
  - Eval = F: Don't evaluate result
- Put R code in text:
  - ex. Write:
    - This is an example `r 1+1`.
  - It will be rendered as:
    - This is and example 2.