Contents

Day 15

Z-Score Example

Two tests of "English" ability

- NAEP Reading Test
- SAT verbal

Suppose a student score scored 320 on NAEP & 650 SAT

Which test did he do better on?

$NAEP \sim N(288, 38)$

$SAT \sim N(500, 120)$

Convert to Z-Scores

NAEP:

$$Z = \frac{value\ -\ mean}{standard\, deviation} = \frac{320 - 288}{38} = 0.842$$

Student scored 0.842 standard deviation above average

SAT:

$$Z = \frac{value\ -\ mean}{standard\, deviation} = \frac{650 - 500}{120} = 1.25$$

Student scored 1.25 standard deviation above average

R-Code

```
pnorm(320, mean = 288, sd = 38)
[1] 0.8001355
```

Cumulative proportion of 0.800 (80%) which means 80$^{\text{th}}$ percentile.

```
pnorm(650, mean = 500, sd = 120)
[1] 0.8943502
```

Cumulative proportion of 0.8943502 which means 89$^{\text{th}}$ percentile.

Water bottle example

Questions

- Why does it continue to overfill
  - How much does it actually pour $\rightarrow$ average
- Why does Dr. Wynne have such terrible reaction speed?
  - Reaction speed $\rightarrow$ average
- Does the water fill at the same rate
  - Average rate for one pour
  - $\rightarrow$ average over several attempts

Expected value = μ = expected amount filled

$\bar{X}$ = average amount filled in a sample of pours "sample mean".

<u>Variability:</u> how variable are the individual values. (range)

- σ = Standard Deviation
- σ^2 = Variance
- S = Sample Standard Deviation
- S^2 = Sample Variance

<u>Bias:</u> <u>Center:</u> - on average, are we where we expected to be? (mean, median, mode)

Shape

Shape: where "average" is compared to "most likely"

- How "consistent" the values are given variability

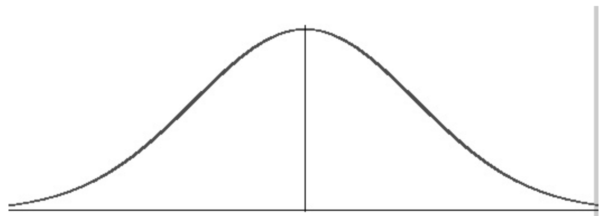

Figure 1: Unimodal Distribution

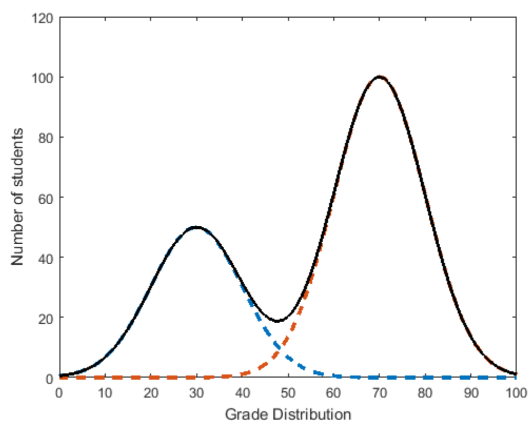

Figure 2: Bimodal Distribution

The median is resistant, the mean is subject to more change.

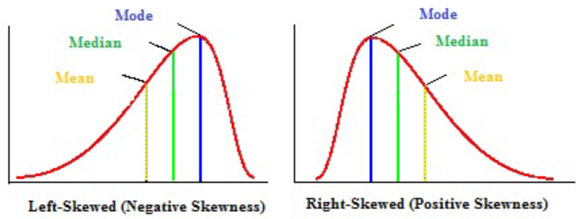

Figure 3: Left and Right Skewed Graphs

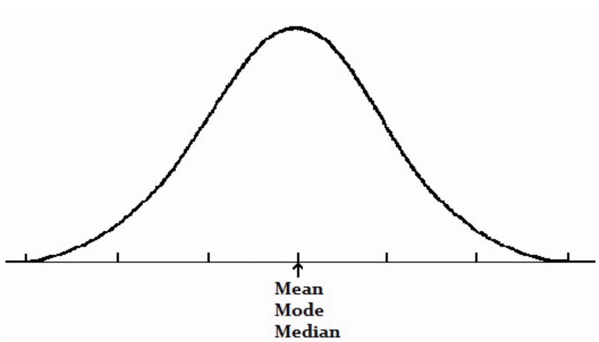

Figure 4: Symmetric Graph

Approximating a Density Curve: Histogram

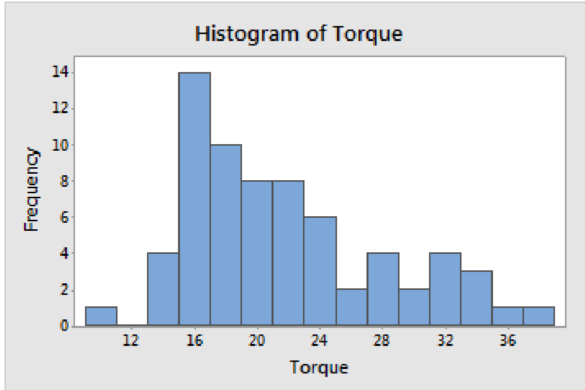

Figure 5: Histogram

- "bins": intervals on the x-axis
- Choice of bins is very important
  - Endpoints of bins
  - Center & Width
- Reimann Integral of an unknown density curve

Outliers

Points that doesn't fit with everything else

Attempting to determine outliers

- Plot your data & look for points that don't belong
- $\uparrow$ best way
- Investigate why they're different

Box-Plots

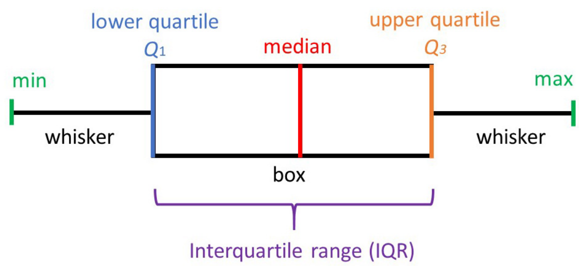

Figure 6: Box Plot

Rule of Thumb

- Step 1: Get five number summary (min, Q_1, medium(Q_2), Q_3, max)
- Step 2: Compute IQR = range middle 50% of data
  - IQR = Q_3 - Q_1
- Step 3: Compute "fences"
  - Lower fence: $Q_1 - K \times IQR$
  - Upper fence: $Q_3 + K \times IQR$

Anything outside the fences is an outlier.

By convention, $k = 1.5$

Example : Senator Ages

Five number summary:

- Min = 39
- Q_1 = 55.5
- Median = 63
- Q_3 = 69
- Max = 85

$IQR = 69 - 55.5 = 13.5$

Lower fence: $55.5 - (1.5)(13.5) = 35.25$ Upper fence: $69 + (1.5)(13.5) = 89.25$

In this data set we have no outliers because our data falls between the fences.

Numerical Variable Connection to Random Variables

Recall for random variable X

$E(A + Bx) = a + b \times E(x)$

$Var(A + Bx) = b^2 \times Var(x)$

$SD(A + Bx) = |b| \times sd(x)$

Recall for random variable X and Y

$E(Ax + By) = aE(x) + bE(y)$

$Var(Ax + By) = A^2 \times var(x) + B^2 var(y)$

$SD(Ax + By) = \sqrt{A^2 \times var(x) + B^2 \times var(y)}$

All of these rules hold for numerical variables too