Chapter 1 Day 2

Summary

The **normal curve** is a **density curve**. The normal curve has the **normal table** which can be used to look up what **proportion** is up to a point. Data can be converted to a **Z-value** where the normal curve and **table** can be used. Knowing the **standard deviation**, **mean**, and z-value we can work backwards to get the **x-value**.

Terminology

Variance / Standard Deviation

- measure variability around the mean
- if number is small, then the spread of the numbers is small
- the unit of variance is squared and not very useful, so we often take the square root to get standard deviation
  Density Curve
- mathematical model of distribution
- an approximation
- The total area under the curve is equal to 1 or 100%
  Population
- **all** items collected
- use greek letters when analysing
  Sample
- by taking a different sample, we will get different numbers
- use variable names

Variance and Standard Deviation

A simple example:
Data set - 1,1,3,4,6
`mean = (1+1+3+4+6)/5 = 3`
xi - xbar

```
1-3=-2
1-3=-2
3-3= 0  # middle
4-3= 1
6-3= 3
```

Now we square it to get rid of negative. Add the squares up to get the **sum of squares**.
`4+4+0+1+9=18`
Then we divide it by `n-1`. Sum of squares is always an underestimate, so we need to divide by `n-1` instead of n.
`18/(5-1) = 4.5`

5 number summary vs mean and std dev

Can use either.
5 number summary usually better if distribution skewed or with strong outliers.
Mean, std. dev. suitable if distribution is reasonably symmetric and free of outliers

Describing distributions

Plot your data.
Look for patterns and outliers. Can they be explained?
Calculate appropriate numerical summary measures.

The density curve

> Sometimes the pattern of observations is so strong we can approximate it with a *density curve*.

Can interpret the area under the curve. That is the proportion of data between x1 and x2.
The **median** of a density curve is the equal areas point.
The **mean** of a desity curve is the balance point.
Can be the same number for a symmetric density curve.
Use English letters for samples, because the numbers will be different for different samples; they are variables. When using the population, however, we use Greek letters.

Normal Curves

The most important class of density curves.
Symmetric with one peak; bell shape.

68, 95, 99.7 rule

Three symmetrical intervals containing proportions.
68 - 1 standard deviation
95 - 2 standard deviations
99.7 - 3 standard deviations

Z-value / Z standard / standardized value of x

Get distance between mean and point, then divide by the standard deviation.

$Z = (x-\mu)/\sigma$
$N(\mu,\sigma)$

Normal Tables

Table A (in textbook) fives proportino of observations that fall to the left of z standard deviations from the mean.

P for proportion

```
P(-2.15 <= Z <= 1.20)  # Proportion between 2 areas
P(Z <= 1.20) - P(Z <= -2.15)
= 0.8849 - 0.0158
= 0.8691 or 86.91%
```

Standardize all terms!

```
P((240-μ)/σ <= (x-μ)/σ <= (275-μ)/σ)
P((240-250)/15 <= Z <= (275-250)/15)
...
```

If you cannot find a Z-value in the table, it is either 0 or 100%!

Working "backwards"

Given proportion, find value

eg.1) 4% of values for the standard normal distribution are bigger than what value.

```
1-0.04 = 0.9600
```

Check closest value at 0.9600. If equally close, take average
1.75

eg.2) The “most central” 90% of standard normal distribution lies between what values?
A: Use symmetry, `1-0.90 = 0.10` for the tails. `0.10/2` = size of each tail.
Between `-1.645 and 1.645`

Also can go the other way, if you have Z, mean, and SD, but need x

```
Z = (x-μ)/σ
Zσ = x-μ
x = μ+Zσ
```