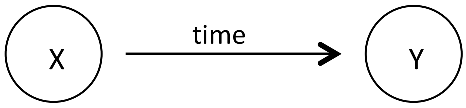
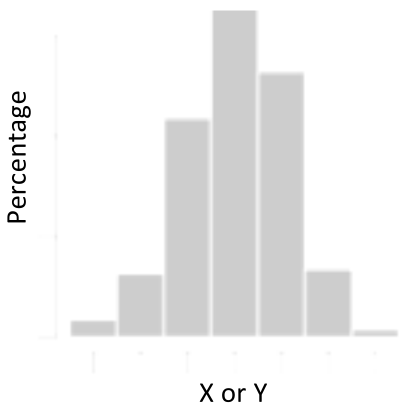
Contents

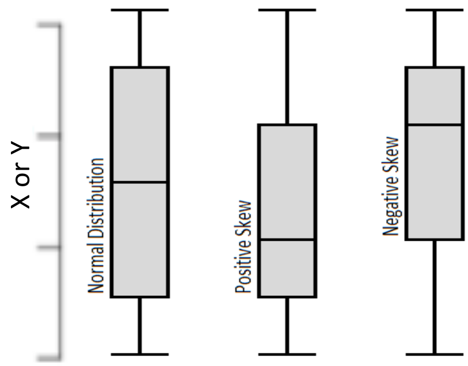
## Session One




### Looking at the individual distributions of X and Y

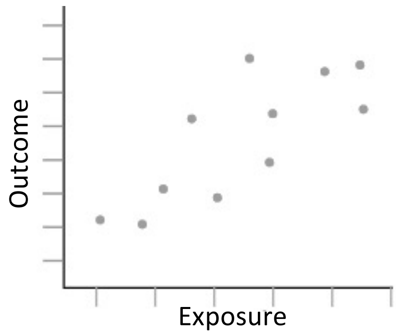
We use either *histograms* or *box plots*




### Looking at the association between X and Y

We use a *scatter plot* for continuous variables




We can comment on:
- **Direction** of the association (+ or –)
- If the association looks **linear** or not
- The **spread** of the data in certain areas

### Simple Linear Regression *reg*

$$\boxed{Y_i = \beta_0 + \beta_i X_i + e_i}$$

intercept slope residual error

- **Continuous outcome**
- Numerical exposure
- Only one exposure variable
- Estimates best-fitting straight line of scatter plot

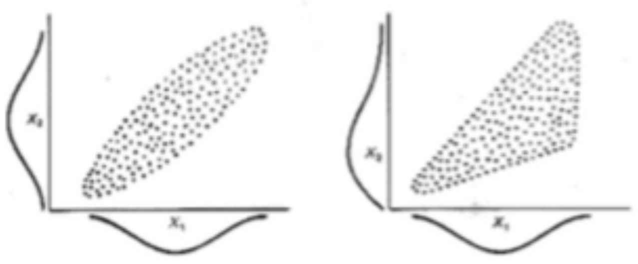
**Coefficient Interpretation:**
**$\beta_0$** = Estimated mean outcome when exposure is zero = `_cons`
**$\beta_1$** = Estimated mean change in outcome for 1 unit increase in exposure

**95% CI interpretation** for $\beta_1$ :
"We are 95% confident that the population mean increase in outcome for a 1 unit increase in exposure could be as low as ## or as high as ##."

**Null hypothesis** of *P*-value for $\beta_1$ :
"There is no association between exposure and outcome; $\beta_1 = 0$"

## Assumptions of linear regression

1. Association between exposure and outcome is **linear**
2. **Residual** variation is **normally distributed**
3. **Independence** of observations
4. **Homoscedasticity** (spread of data points is consistent and does not fan out)



Homoscedasticity with both variables normally distributed    Heteroscedasticity with skewness on one variable

## Methods of least squares

Used to derive the best-fitting line of a simple linear regression model.
It finds values of $\beta_0$ and $\beta_1$ that minimize the sum of the squared vertical distances from observed points to the fitted line.
We square these residuals so that the negatives and positives (observed points above and below the fitted line) do not nullify each other.

## Interchanging exposure and outcome

X → Y    $\neq$    X → Y

The **slope parameter** for Y on X is:

$m_1 \times r^2$    or    $\frac{1}{\beta 1} \times r^2$    … where **r** is the correlation coefficient

## Correlation Coefficient (r)

The correlation coefficient can take values between $-1 \leq r \leq +1$.
It tells you the **direction** (+ or –) and **strength** (closer to zero=weak; away from zero=strong) of association.
It has the same sign (+ or –) as the **regression coefficient** $\beta_1$.

$$\hat{\beta}_1 = r\frac{s_y}{s_x}$$

…where $s_x$ and $s_y$ are the sample standard deviations of x and y.

If $s_x = s_y$, then $\beta_1 = r$

How do we get the **estimated slope parameter** if we have $\boldsymbol{\beta_1}$ (0.0436) and **r** (0.7591)?

$$\text{estimated slope parameter} = \frac{1}{\beta 1} \times r^2$$

$$= \frac{1}{0.0436} \times (0.7591)^2 = 13.22$$

## Centering a variable around a value

We do this to obtain a sensible, more logical interpretation of the intercept ($\beta_0$).

$X_{centred} = X_{original}$ *– new centre*

When we center a variable around a value and fit a regression…

1. the slope parameter will not change (**$\beta_1$**)
2. the value and interpretation of the intercept (**$\beta_0$**) changes
   *"$\beta_0$ is the estimated mean outcome value when the exposure = (new center)"*

## Centering a variable around it's mean

The mean of the new centered variable = 0

$X_{centred} = X_{original}$ *– mean of X*

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Eg. SBP = 125 – 4.7 x Birthweight

Q. The sample mean birthweight is 2.7kg. If we subtract the sample mean from each participant's birthweight to create "$\text{Birthweight}_{\text{adjusted}}$" and then fit a regression line, what would the values of *a* and *b* be below?
SBP = *a* + *b* x $\text{Birthweight}_{\text{adjusted}}$

A. Y = $\beta_0 + \beta_1$ x $\text{Birthweight}_{\text{adjusted}}$
When we centre variables around the mean, which is what this question has done, $\beta_1$ stays the same but $\beta_0$ changes. Now, $\beta_0$ is the estimated mean SBP when the exposure = 2.7kg. The original model we were given was:
SBP = 125 – 4.7 x Birthweight …so:
Predicted SBP = 125 – 4.7 x 2.7
Predicted SBP = 112.31mmHg = $\beta_0$ ...so:
$\beta_0$ = *a* = 112.31 mmHg
$\beta_1$ = *b* = -4.7

## Standardized variables

When we standardize the exposure *and* outcome variables and fit a linear regression, the **slope parameter ($\beta_1$)** will be the **correlation coefficient** of the two variables.

$$z = \frac{x - \mu}{\sigma}$$

The correlation coefficient of the standardized variables is still the same coefficient of the original variables.