# Applied Statistical Analysis I: Bivariate Regression and Topic Review

Dr Redmond Scales

Trinity College Dublin
rscales@tcd.ie

October 21, 2025

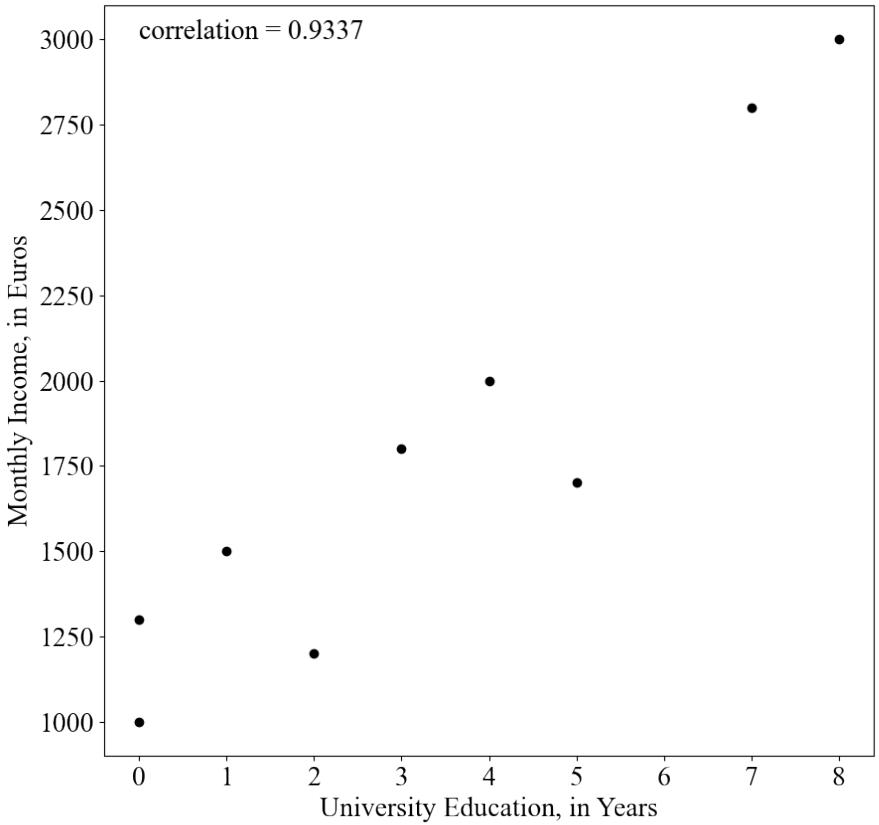
# Today's Agenda

(1) Lecture recap & exam review
(2) Git pull
(3) Tutorial exercises

# Linear regression model

*What is a linear regression model? What interpretations can we make?*

# So far, correlation analysis



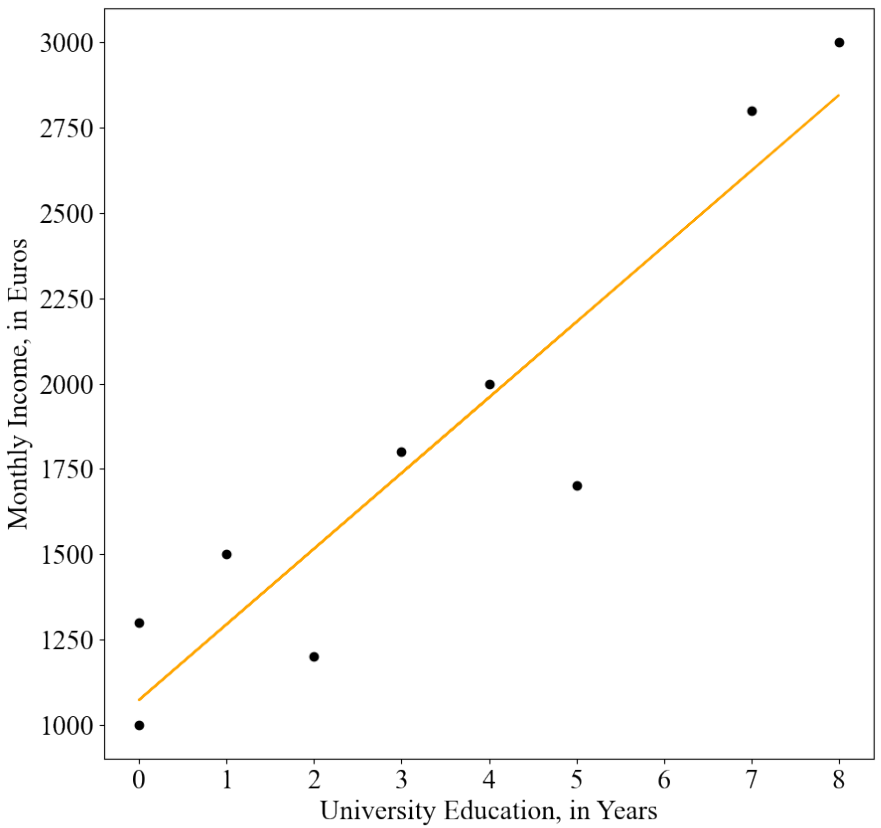

Just by looking at the plot, can you identify the straight line which best describes the joint variation between $X$ and $Y$?

*This is fictional data.

# Regression analysis

*What is a linear regression model?*

- Find linear line of best fit, $Y_i = \alpha + \beta X_i + \epsilon_i$



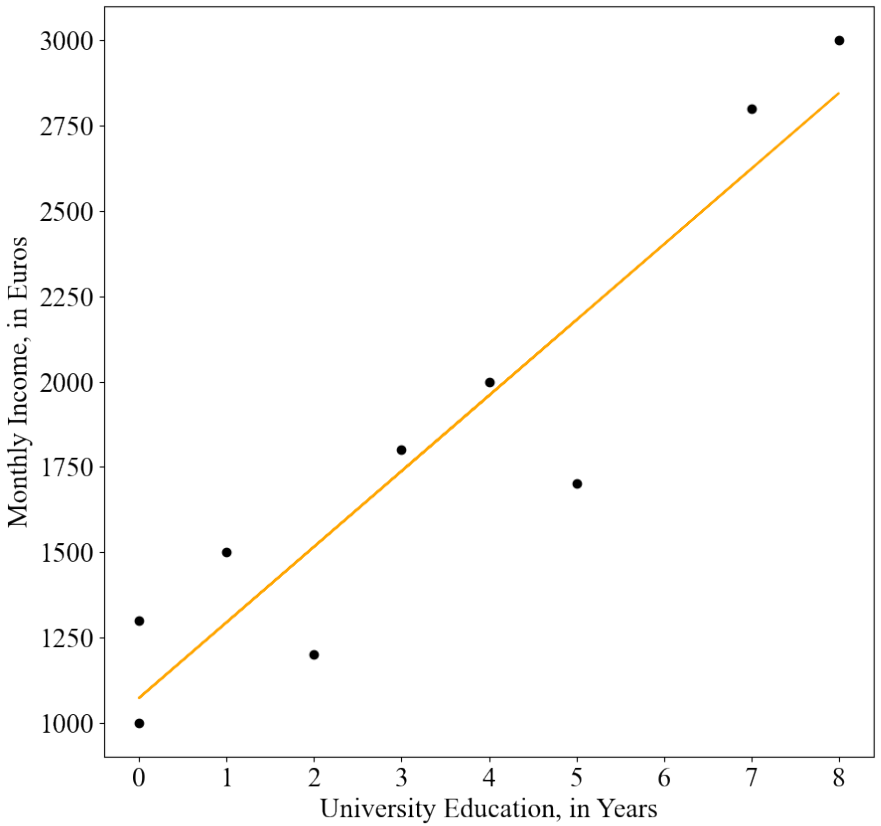
# Regression analysis

*What is a linear regression model?*

- Find linear line of best fit, $Y_i = \alpha + \beta X_i + \epsilon_i$
- $\alpha$ (intercept): expected value of $Y$ when $X = 0$
- $\beta$ (slope): expected change in $Y$ when $X$ increases by one unit
- $\hat{Y}$ (expected value): predicted outcome based on the regression model, $\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$
- $\epsilon$ (error/residual): difference between actual and predicted outcome, $\epsilon_i = Y_i - \hat{Y}_i$

# Regression analysis

*What interpretations can we make?*

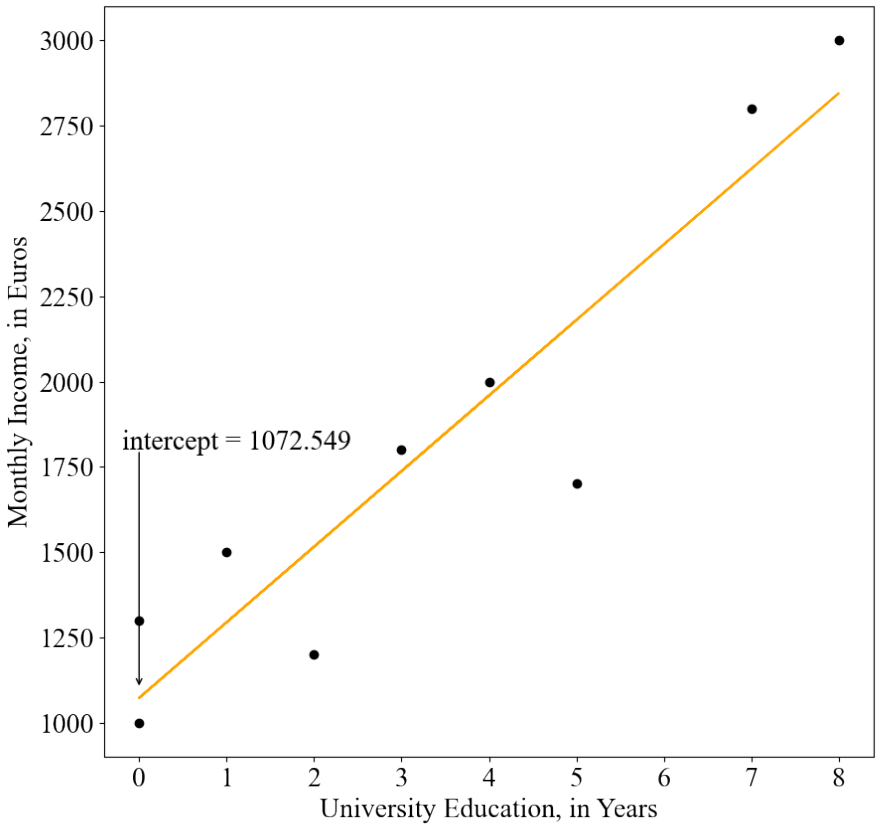



$$income = \alpha + \beta * education$$

$$income = 1072.5490 + 221.5686 * education$$

# Regression analysis

*What interpretations can we make?* (intercept)

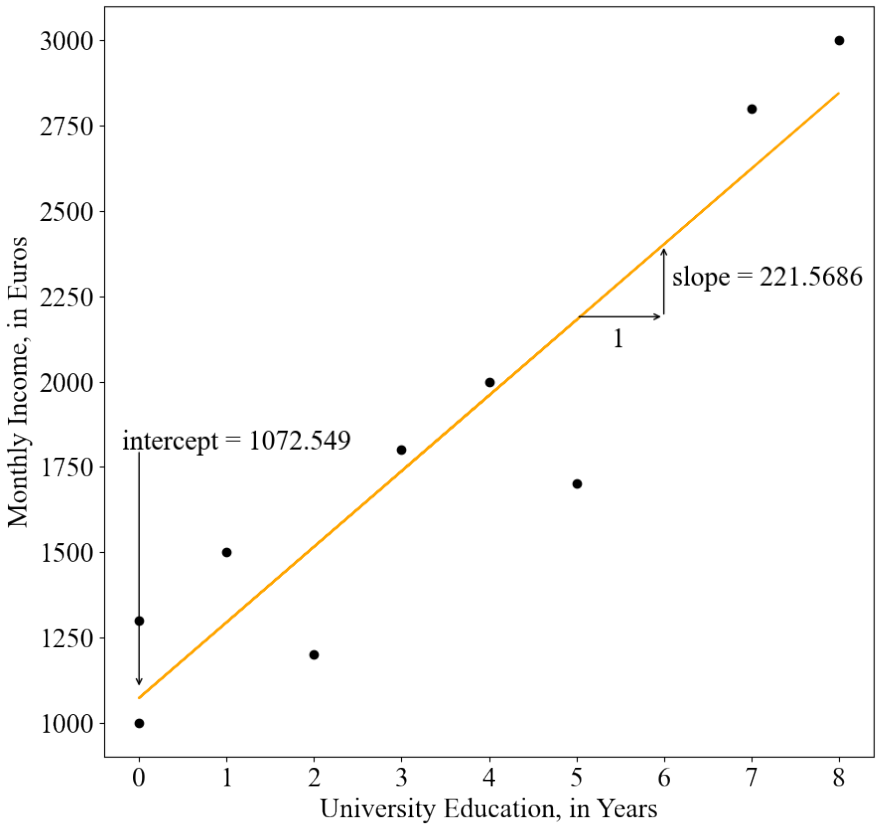



If an individual has a university education of 0 years, what income would we expect for that person?

$$income = 1072.5490 + 221.5686 * 0 = 1072.5490$$

# Regression analysis

*What interpretations can we make?* (slope)

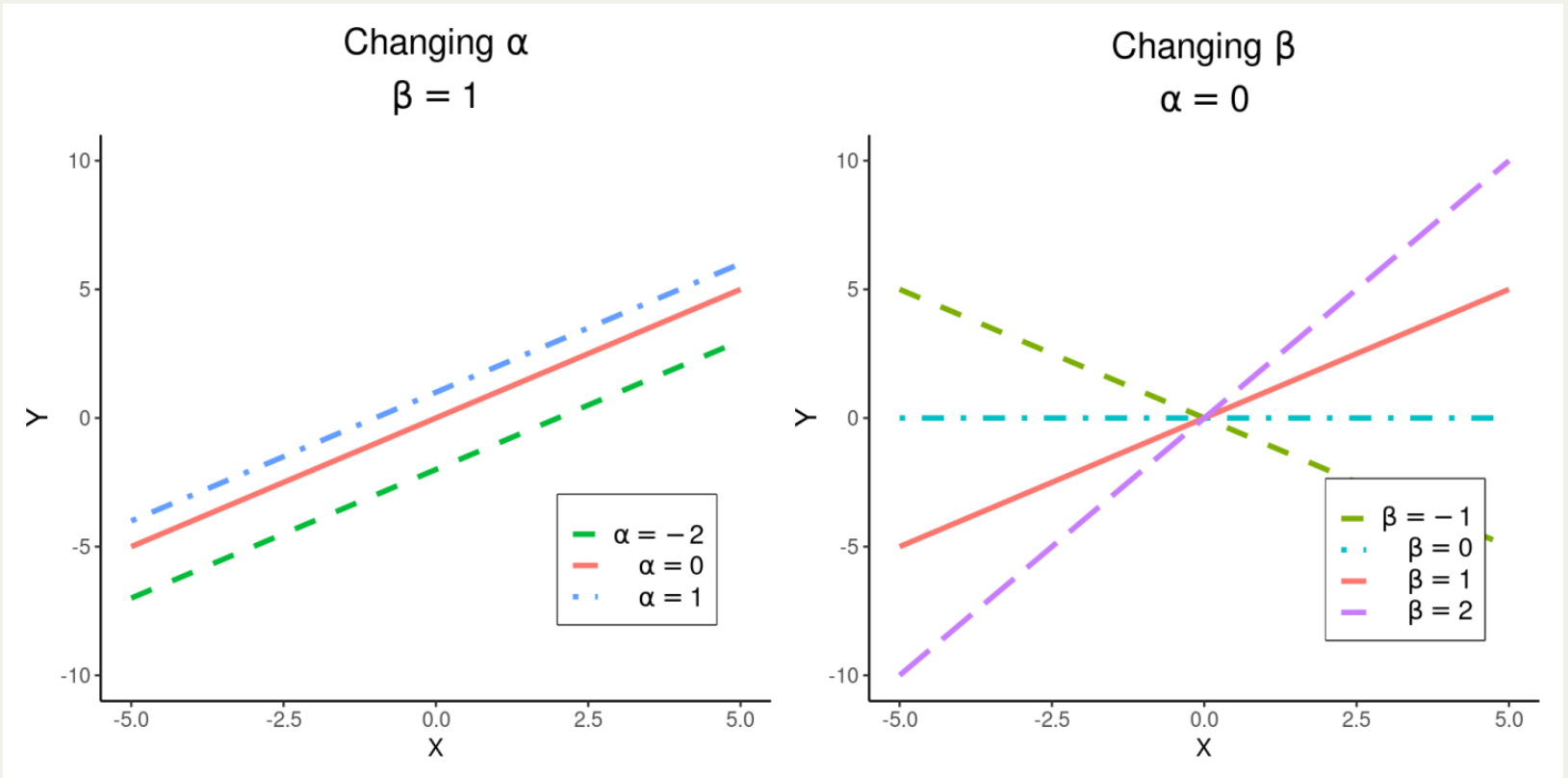



If the university education increases by one year, how much more Euros would we expect an individual to earn? $income = 1072.5490 + 221.5686 * 1 = 1294.1176$

$\rightarrow$ With every additional year of university education, the expected income increases by 221.5686 Euros.

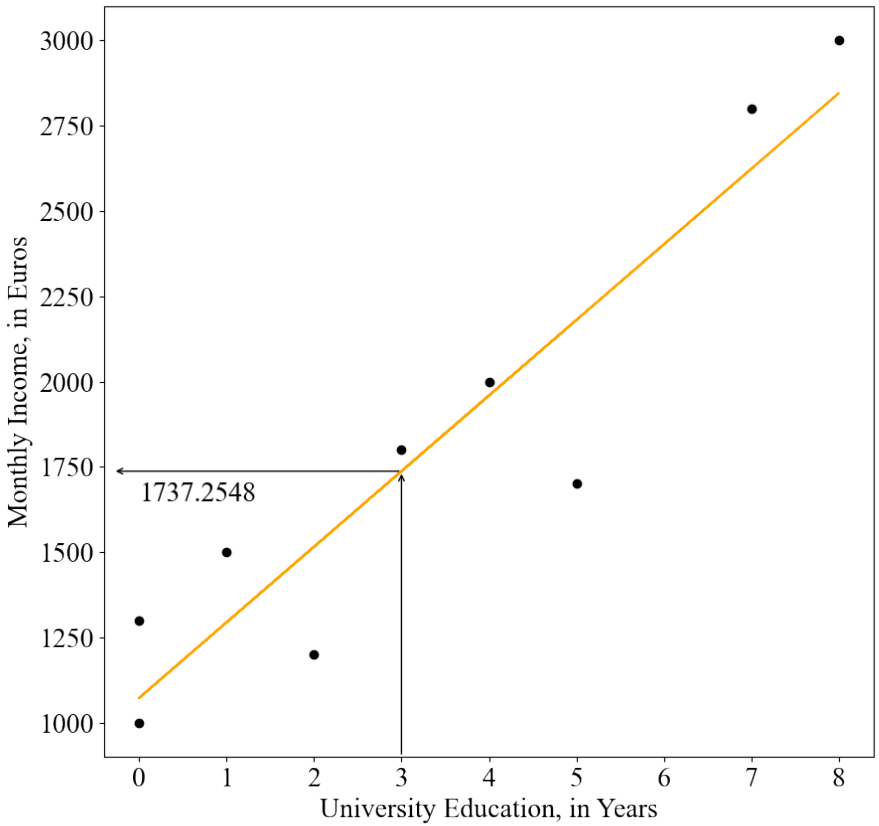
# Regression analysis

## Varieties of linear relationships



# Regression analysis

*What interpretations can we make?* (expected value)

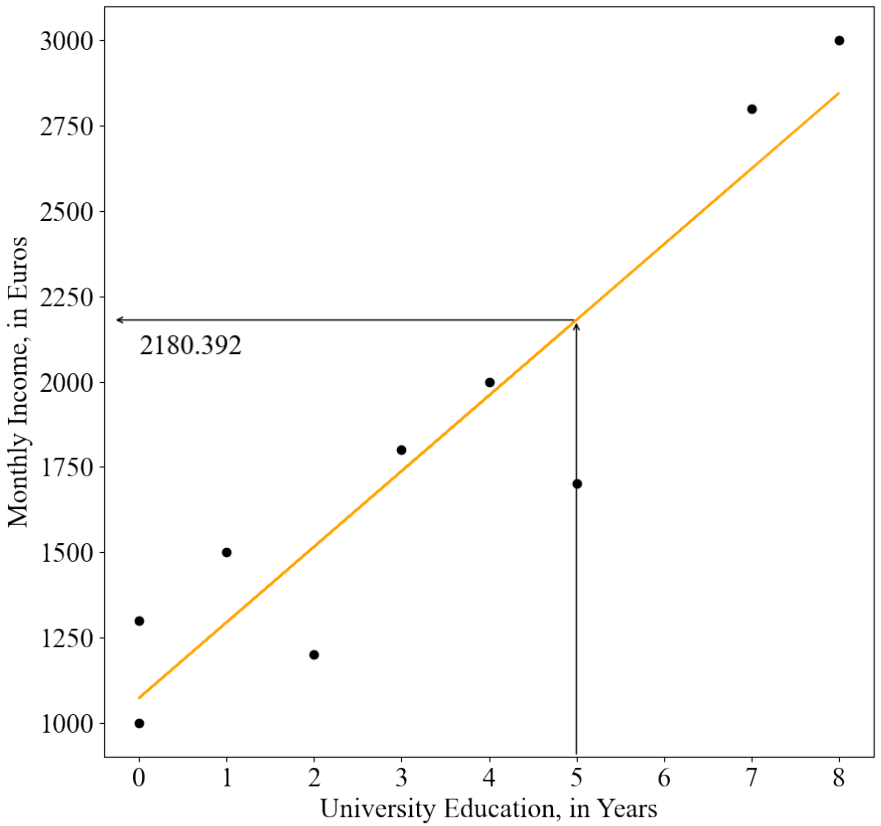



If an individual has 3 university education years, what income would we expect for that person?

$$income = 1072.5490 + 221.5686 * 3 = 1737.2548$$

# Regression analysis

*What interpretations can we make?* (residual)

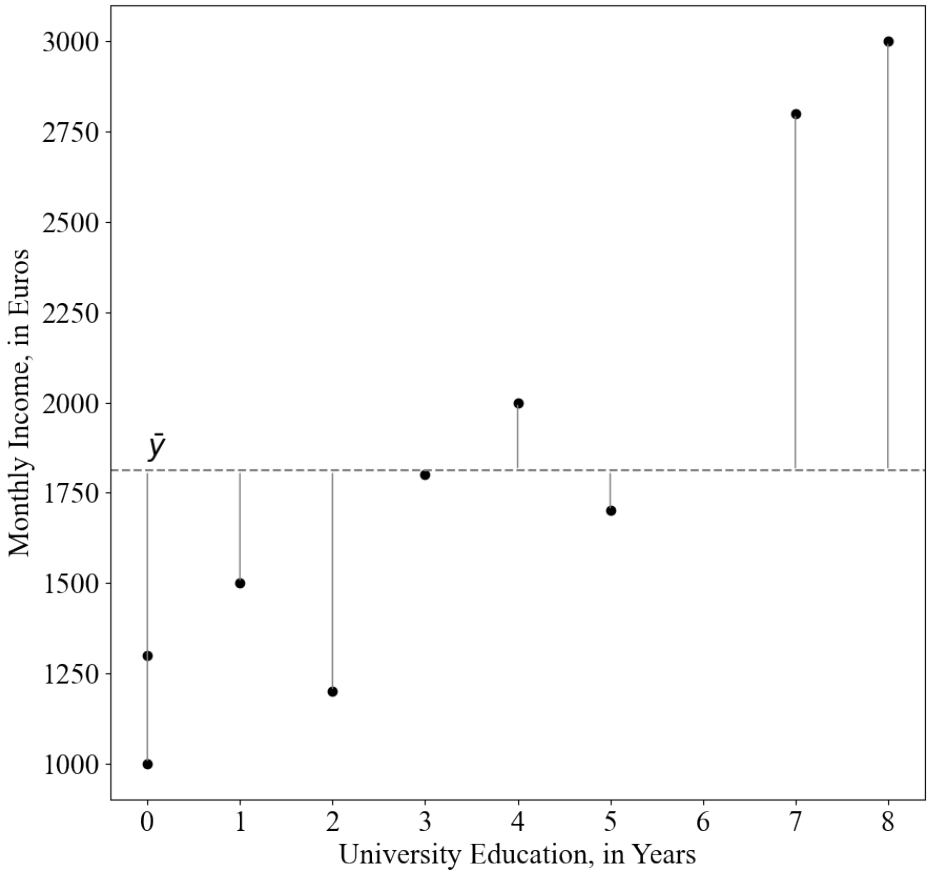



$income = 1072.5490 + 221.5686 * 5 = 2180.392$

Residual = Actual – Predicted

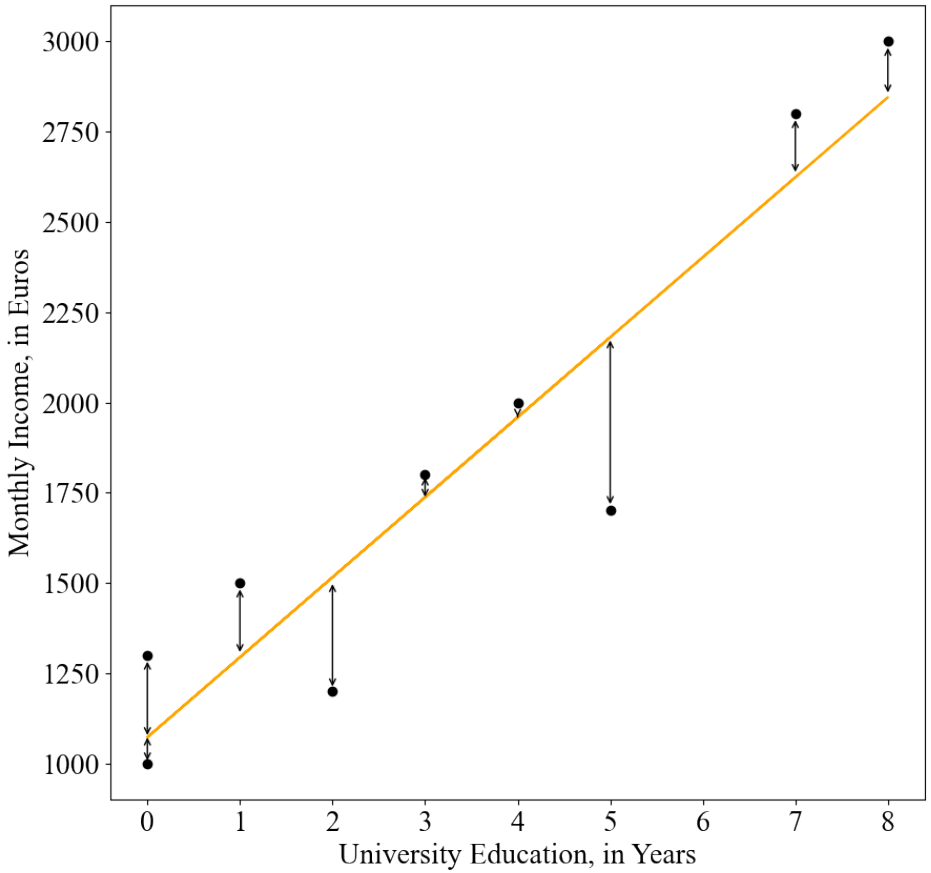
Residual = 1700 – 2180.392 = –480.392

# Total sum of squares ($SS_T$)



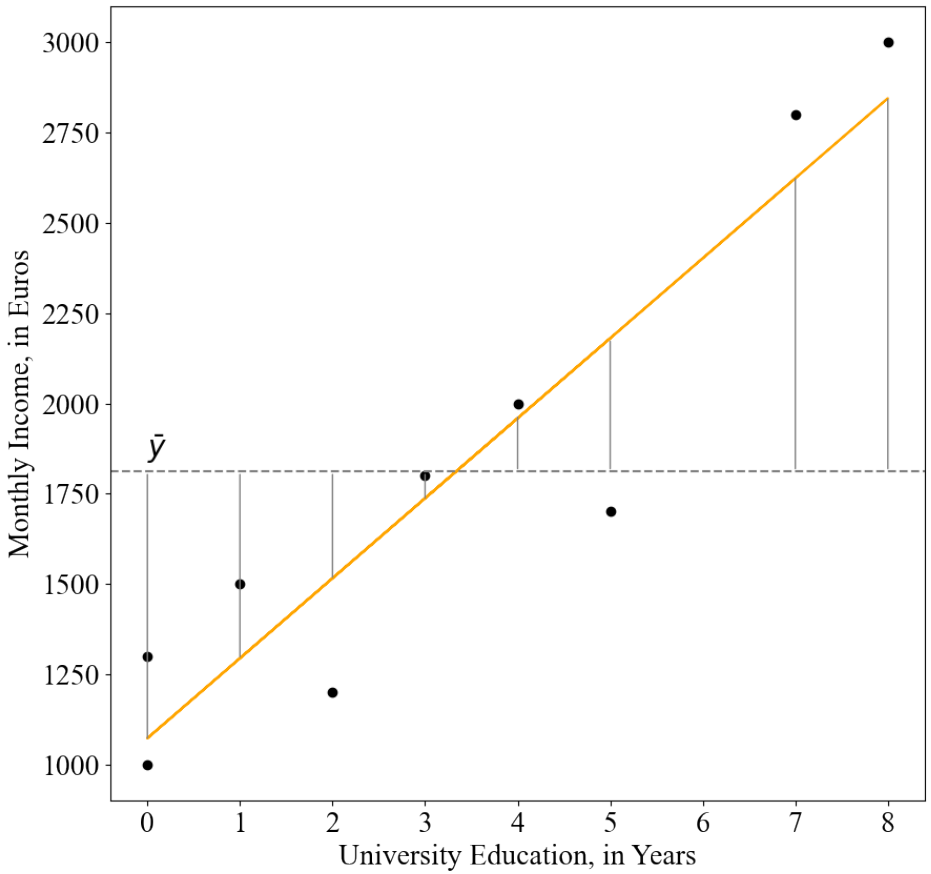

$SS_T$ = Sum of squared differences between observed values of Y and the mean, $SS_T = \sum_{i=n}^{n}(y_i - \bar{y})^2$

# Residual sum of squares ($SS_R$)




$SS_R$ = Sum of squared differences between observed values of Y and the regression line, $SS_R = \sum_{i=n}^{n}(y_i - \hat{y}_i)^2$

# Model sum of squares ($SS_M$)




$SS_M$ = Sum of squared differences between the regression line and the mean of Y, $SS_M = \sum_{i=n}^{n}(\hat{y}_i - \bar{y})^2$

$\rightarrow$ Improvement if regression model is used rather than the mean.

# Regression analysis

*What interpretations can we make?* (model performance)

- $R^2$: the proportion of variation of $Y$ explained by $X$. Varies between 0 and 1. If $X$ explains all the variation in $Y$, then $R^2 = 1$.

$$SS_T = SS_M + SS_R \text{ and } SS_M = SS_T - SS_R$$

$$R^2 = 1 - \frac{SS_R}{SS_T} =$$

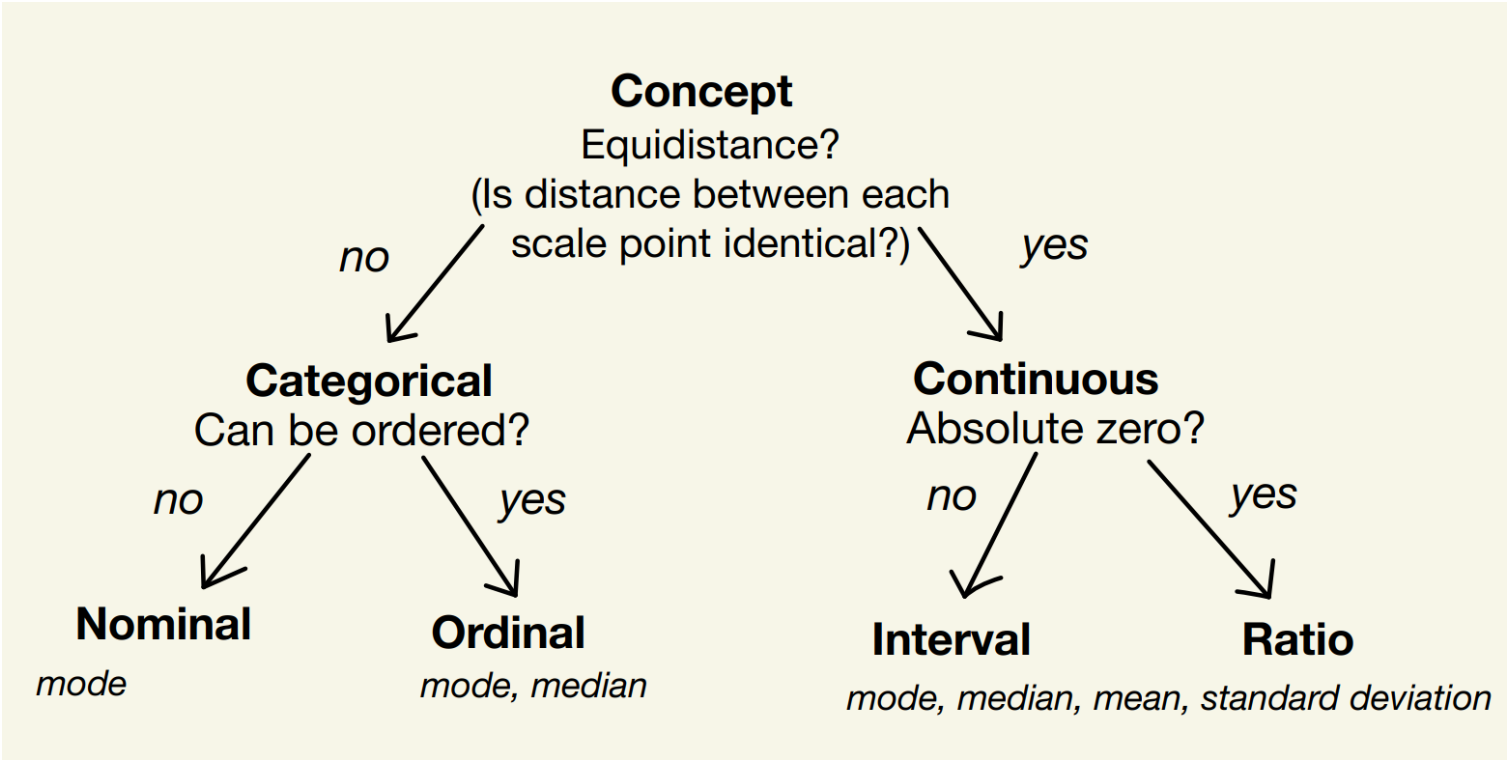
$$1 - \frac{\text{Variation not explained by model}}{\text{Total variation in y}} =$$

$$\frac{SS_M}{SS_T} = \frac{\text{Variation explained by model}}{\text{Total variation in y}}$$

*Week 1–Introduction & stats review*

# Measurement Scales

*How can we measure concepts? And why does it matter?*




(Kellstedt and Whitten 2018, Chap. 5)

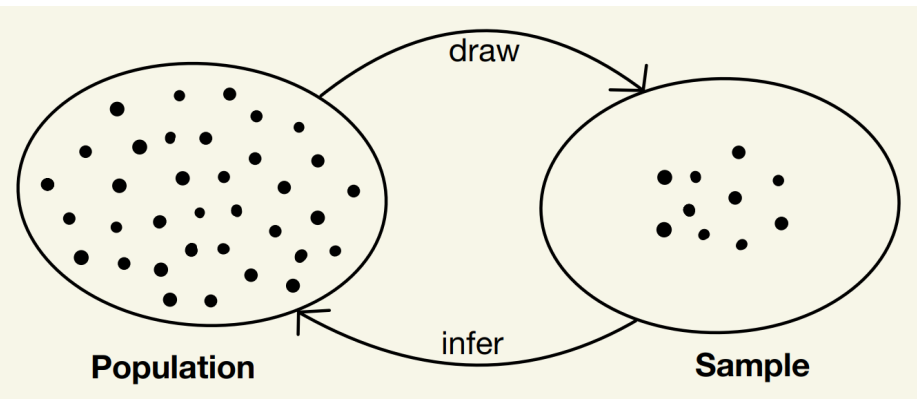
Discrete: finite set of possible values.

Continuous: infinite set of possible values.

# Population, sample, parameter, statistic

*What is the relationship between population and sample?*

- Population: "the total set of subjects of interest in a study" (Agresti and Finlay 2009, 5).
- Parameter: "numerical summary of the population" (Agresti and Finlay 2009, 5).
- Sample: "the subset of the population on which the study collects data" (Agresti and Finlay 2009, 5).
- Statistic: "a numerical summary of the sample data" (Agresti and Finlay 2009, 5).
- Observation: single subject/unit, one row in dataset



# Measures of central tendency and variability (dispersion)

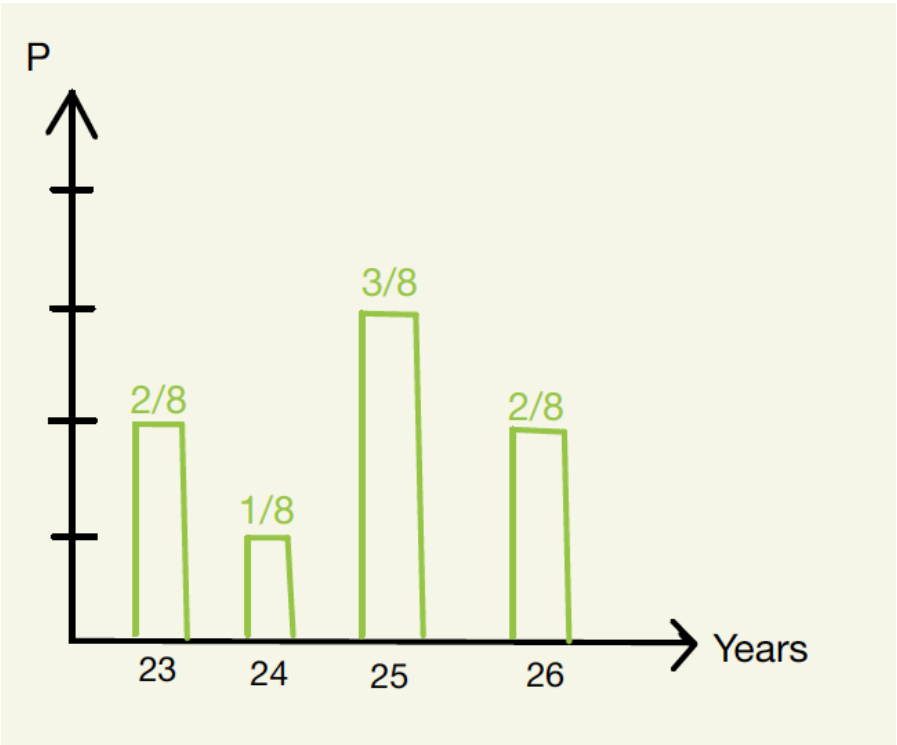
*How can we describe variables?*

- Mean: $\bar{y}$ = Sum of all values divided by the number of observations, $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$
- Variance: $s^2(y)$ = Sum of squared deviations divided by number of observations (deviation is the difference between observed value and the mean, $y_i - \bar{y}$), $s^2 = \frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n-1}$
- Standard Deviation: Return original units by taking square root, $s = \sqrt{\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n-1}}$

# Distributions and probability distributions

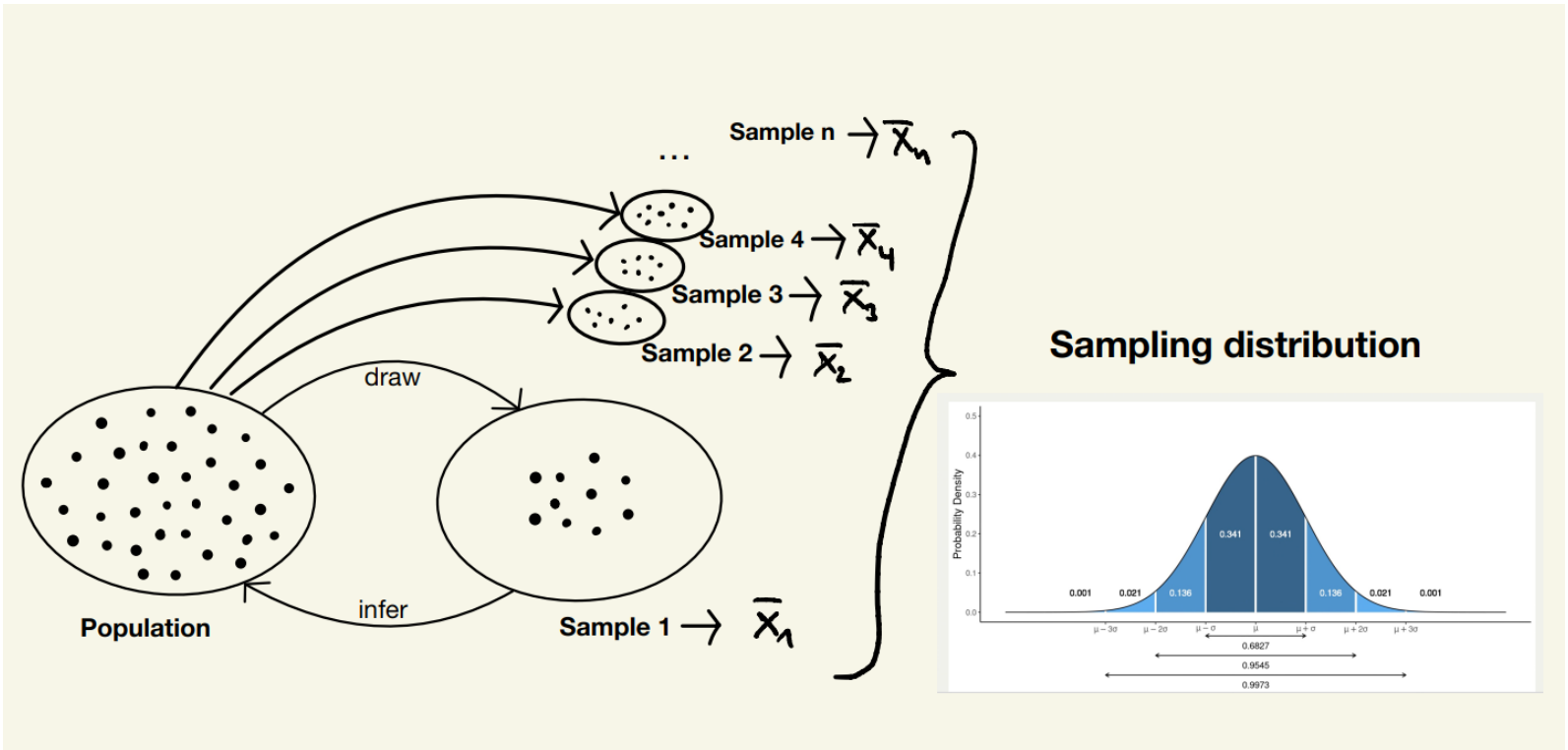
*What is a probability distribution?*

- Probability distribution "lists the possible outcomes and their probabilities" (Agresti and Finlay 2009, 75).



# Sampling distribution

theoretically...



# Sampling distribution

*What is a sampling distribution?*

- Sampling distribution "A sampling distribution of a statistic is the probability distribution that specifies probabilities for the possible values the statistic can take" (Agresti and Finlay 2009, 87).
- In other words, a probability distribution for a statistic rather than values of observations $\rightarrow$ What is the probability of $\bar{Y} = 0.5$, rather than what is the probability of $Y = 3$?

# Sampling distribution

*Why is this important?*

- The corresponding probability theory "helps us predict how close a statistic falls to the parameter it estimates" (Agresti and Finlay 2009, 87). $\rightarrow$ how close is $\bar{y}$ to $\mu$?
- Usually only one sample/one estimate $\rightarrow$ Point estimate: "is a single number that is the best guess for the parameter value" (Agresti and Finlay 2009, 107).

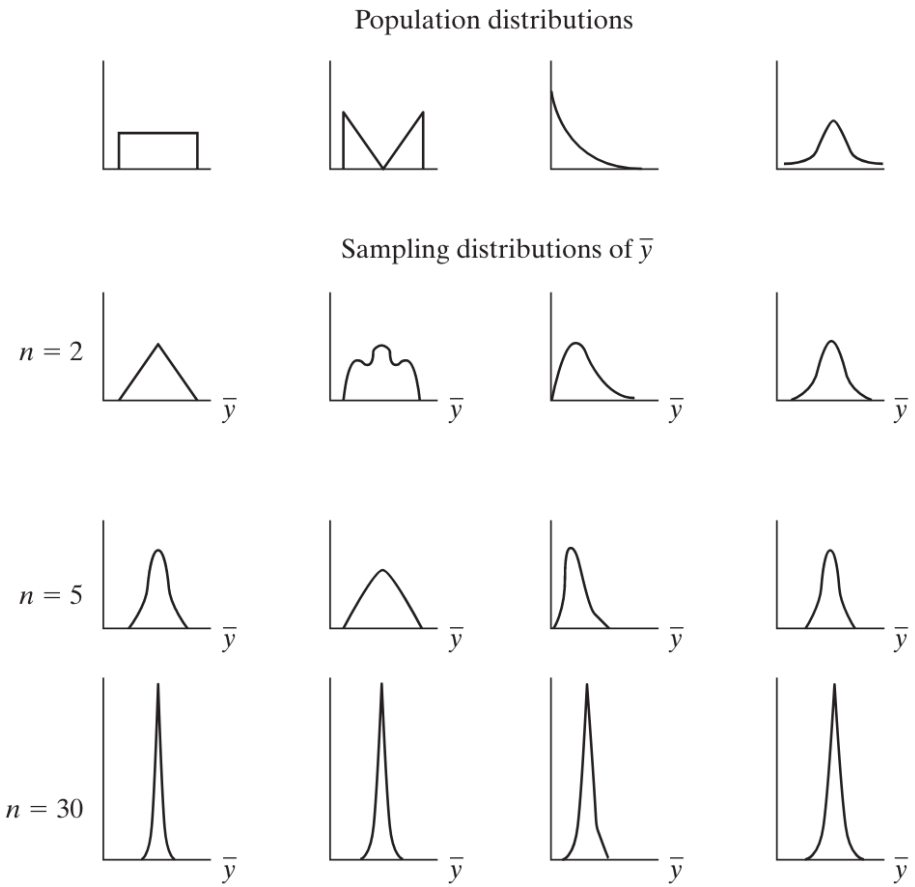
# The sampling distribution of the mean, $\bar{y}$

- "If we repeatedly took samples, then in the long run, the mean of the sample means would equal the population mean $\mu$" (Agresti and Finlay 2009, 90). $\rightarrow$ mean of the sampling distribution of $\bar{y}$ equals the population mean, hence, $\mu = \bar{y}$
- "The standard error describes how much $\bar{y}$ varies from sample to sample" (Agresti and Finlay 2009, 90). $\rightarrow$ standard error is estimated based on standard deviation, hence, $\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}}$
- *Why does this work?*

# Central Limit Theorem

*What is the Central Limit Theorem?*

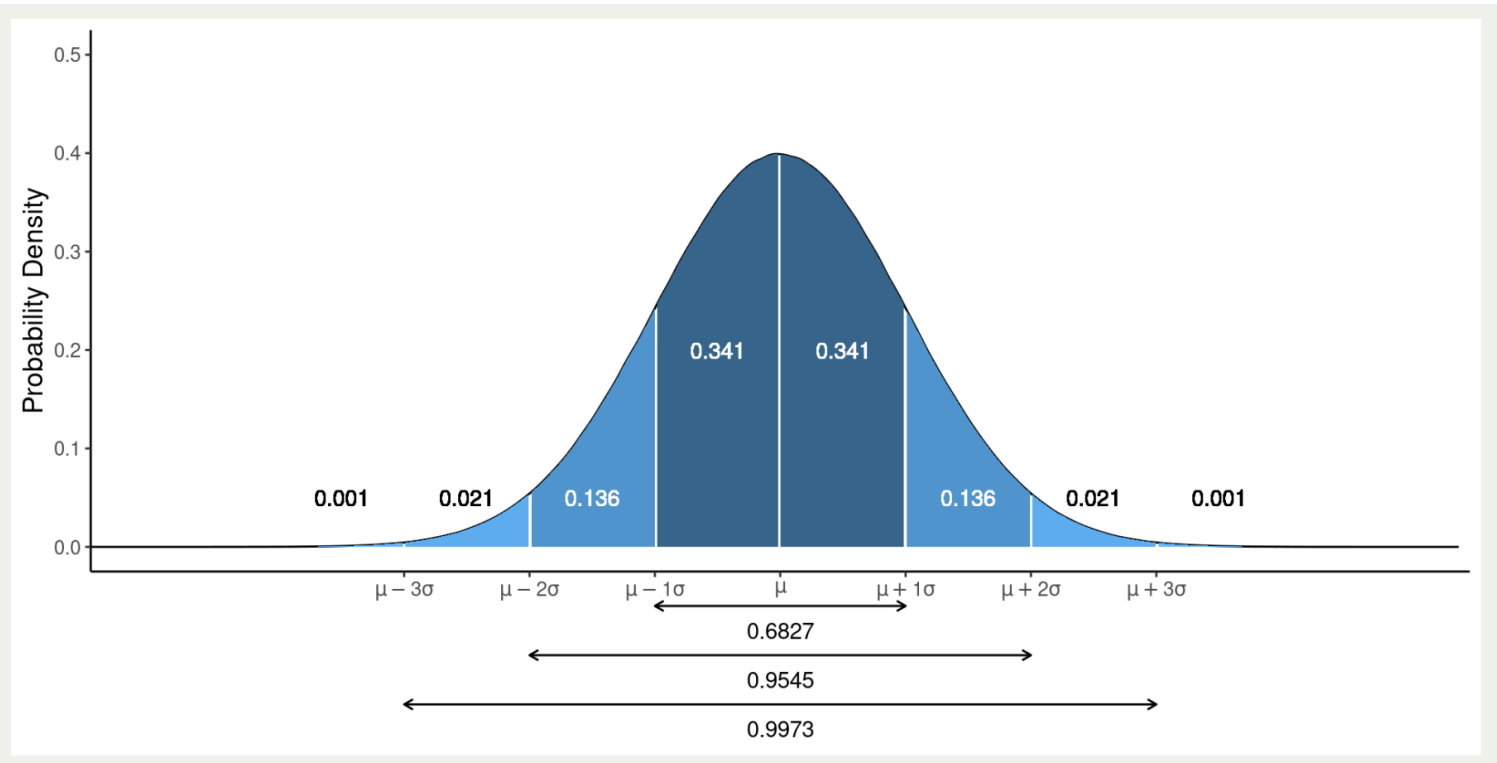
- "For random sampling with a large sample size $n$, the sampling distribution of the sample mean $\bar{y}$ is approximately a normal distribution" (Agresti and Finlay 2009, 93). $\rightarrow$ regardless of the population distribution



# Central Limit Theorem

*What is the Central Limit Theorem?*

- "Knowing that the sampling distribution of $\bar{y}$ can be approximated by a normal distribution helps us to find probabilities for possible values of $\bar{y}$ (Agresti and Finlay 2009, 94). $\rightarrow$ key in inferential statistics



# Confidence intervals

*What are confidence intervals?*

- Confidence interval: "an interval of numbers around the point estimate that we believe contains the parameter value" (Agresti and Finlay 2009, 110). $\rightarrow$ Point estimate $\pm$ Margin of error
- Confidence level: "The probability that this method produces an interval that contains the parameter" (usually 0.95, 0.99) (Agresti and Finlay 2009, 110).
- Margin of error = multiple of the standard error, $\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}}$ (Agresti and Finlay 2009, 117).
- For example, for 95% confidence level, the margin of error is $\pm 1.96\sigma_{\bar{y}}$ (have a look at the normal distribution).

*Week 2–Hypothesis testing, experiments, difference in means*

# Null-hypothesis significance testing

**TABLE 6.1:** The Five Parts of a Statistical Significance Test

1. **Assumptions**
   Type of data, randomization, population distribution, sample size condition
2. **Hypotheses**
   Null hypothesis, $H_0$ (parameter value for "no effect")
   Alternative hypothesis, $H_a$ (alternative parameter values)
3. **Test statistic**
   Compares point estimate to $H_0$ parameter value
4. ***P*-value**
   Weight of evidence against $H_0$; smaller $P$ is stronger evidence
5. **Conclusion**
   Report $P$-value
   Formal decision (optional; see Section 6.4)

(Agresti and Finlay 2009, 147)

# Significance test for a mean (t-test)

**TABLE 6.3:** The Five Parts of Significance Tests for Population Means

1. **Assumptions**
   Quantitative variable
   Randomization
   Normal population (robust, especially for two-sided $H_a$, large $n$)
2. **Hypotheses**
   $H_0$: $\mu = \mu_0$
   $H_a$: $\mu \neq \mu_0$ (or $H_a$: $\mu > \mu_0$ or $H_a$: $\mu < \mu_0$)
3. **Test statistic**
   $t = \dfrac{\bar{y} - \mu_0}{se}$ where $se = \dfrac{s}{\sqrt{n}}$
4. ***P*-value**
   In $t$ curve, use
   $P$ = Two-tail probability for $H_a$: $\mu \neq \mu_0$
   $P$ = Probability to right of observed $t$-value for $H_a$: $\mu > \mu_0$
   $P$ = Probability to left of observed $t$-value for $H_a$: $\mu < \mu_0$
5. **Conclusion**
   Report $P$-value. Smaller $P$ provides stronger evidence against $H_0$ and supporting $H_a$. Can reject $H_0$ if $P \leq \alpha$-level.

(Agresti and Finlay 2009, 155)

# Significance test for a difference in means (t-test)

*What is a t-test for the difference in means?*

- Null and alternative hypothesis: (Step 2) The means of two groups are identical, $\bar{y}_1 = \bar{y}_2$ or $\bar{y}_1 - \bar{y}_2 = 0$ ($H_0$), the means of two groups are different, $\bar{y}_1 \neq \bar{y}_2$ ($H_a$).
- Test statistics: (Step 3) "measures the number of standard errors between the estimate and the $H_0$ value" (Agresti and Finlay 2009, 192).

$$t = \frac{\textit{Estimate of parameter} - \textit{Null hypothesis value of parameter}}{\textit{Standard error of estimate}}$$

$t = \frac{(\bar{y}_1 - \bar{y}_2) - 0}{se}$, $H_0$ assumes $\bar{y}_2 - \bar{y}_1 = 0$, $se = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

# Causal effect

*What is a causal effect?*

- Causal effect: "change in some feature of the world that would result from a change to some other feature of the world", $Y_{T=1,i} - Y_{T=0,i} = Y_i^1 - Y_i^0$
- Counterfactual comparison: "outcome would be different in a counterfactual world in which the action was different" $\rightarrow$ what would be the state of $Y$, had $X$ not occurred?
- Fundamental problem of causal inference: "we can only observe, at most, one of the two quantities–$Y_{1i}$ or $Y_{0i}$–for any individual at a particular point in time" (Bueno de Mesquita and Fowler 2021, 164). $\rightarrow$ causal effect is unobservable

(Bueno de Mesquita and Fowler 2021, 159)

# Sample average treatment effect, Difference in means

- Sample average treatment effect (SATE) is unobservable due to fundamental problem of causal inference $\rightarrow$ we only observe sample difference in means
- Sample difference in means is biased estimate of the true SATE

  $\rightarrow$ **Correlation does not imply causation**
- Baseline differences: "[d]ifference in the average potential outcome between two groups (e.g., the treated and untreated groups), even when those two groups have the same treatment status" $\rightarrow$ Confounders may cause baseline differences, which may cause bias (*omitted variable bias*)

(Bueno de Mesquita and Fowler 2021, 187)

*Week 3–Contingency tables, correlation & bivariate regression*

*Week 4–Bivariate regression, inference & prediction*

# Chi-square test of independence

*What is the Chi-square test of independence?*

- Null and alternative hypothesis: Two variables are independent, $f_o = f_e$ ($H_0$), two variables are dependent, $f_o \neq f_e$ ($H_a$).
- Test statistics: "compares the observed frequencies in the contingency table with values that satisfy the null hypothesis of independence" (Agresti and Finlay 2009, 225), $X^2 = \sum \frac{(f_o - f_e)^2}{f_e}$

# Correlation

*How can we measure correlation?*

- Covariance: covariance is the average of the product of deviations of two quantitative variables from the mean, $cov(X, Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n}$ (only interpret sign)
- Correlation: (correlation coefficient, Pearson correlation coefficient, Pearson's $r$, $r$) standardized average of the product of deviations of two variables from the mean (=standardized covariance), $r_{xy} = \frac{covariance(XY)}{S_X S_Y}$ (interpret magnitude, range -1 and 1)

# Correlation

*How can we test the statistical significance of correlation?*

- Null and alternative hypotheses:
  - there is no association between $X$ and $Y$, $\rho_{xy} = 0$ ($H_0$)
  - there is an association between $X$ and $Y$, $\rho_{xy} \neq 0$ ($H_a$)
- Test statistic: $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$ (in $R$)
- Test statistic: $t = \frac{r}{\sqrt{1-r^2/n-2}}$ (in Agresti and Finlay 2009)

# Ordinary least squares (OLS)

*How are intercept and slope estimated?*

- How do we find the line which best fits the data?
- Apply the OLS (Ordinary Least Squares) method, which minimizes the sum of squared errors (SSE).
- Sum of squared errors = the sum of squared differences between actual and predicted values of $Y$.
- $SSE = \sum_{i=1}^{n}(\hat{\epsilon}_i)^2 = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n}(Y_i - (\hat{\alpha} - \hat{\beta}X_i))^2$ $\rightarrow$ minimize this!

# Assumptions of linear regression

Assumptions about the error ($\epsilon_i$), $Y_i = \alpha + \beta X_i + \epsilon_i$

$$\epsilon_i \sim N(0, \sigma^2)$$

* $\epsilon_i$ is normally distributed $\rightarrow$ needed for inference
* $E(\epsilon_i) = 0$, no bias $\rightarrow$ violated if error is not random, but correlated with omitted variable
* $\epsilon_i$ has constant variance $\sigma^2$ (Homoscedasticity $\leftrightarrow$ Heteroscedasticity)
* No autocorrelation, "Autocorrelation occurs when the stochastic terms for any two or more cases are systematically related to each other".
* X values are measured without error

(Kellstedt and Whitten 2018, 190–194)

# Assumptions of linear regression

Assumptions about the model specification, $Y_i = \alpha + \beta X_i + \epsilon_i$

* No causal variables left out and no noncausal variables included
* Parametric linearity

(Kellstedt and Whitten 2018, 190–194)

# Assumptions of linear regression

Minimal mathematical requirements, $Y_i = \alpha + \beta X_i + \epsilon_i$

* X must vary
* Number of observations must be larger than the number of predictors
* In multiple regression: No perfect multicollinearity

(Kellstedt and Whitten 2018, 190–194)

# Inference about the slope

*What is the t-test for the slope of a regression line?*

- Null and alternative hypotheses:
  - there is no association between $X$ and $Y$, $\beta = 0$ ($H_0$)
  - there is an association between $X$ and $Y$, $\beta \neq 0$ ($H_a$)
- Test statistic: "measures the number of standard errors between the estimate and the $H_0$ value" (Agresti and Finlay 2009, 192).

$$t = \frac{\textit{Estimate of parameter} - \textit{Null hypothesis value of parameter}}{\textit{Standard error of estimate}}$$

$$t = \frac{\hat{\beta} - \beta_{H_0}}{se_{\hat{\beta}}} = \frac{\hat{\beta}}{se_{\hat{\beta}}}, \; H_0 \text{ assumes } \beta = 0$$

# References I


Agresti, Alan, and Barbara Finlay. 2009. *Statistical methods for the social sciences.* Essex: Pearson Prentice Hall.

Bueno de Mesquita, Ethan, and Anthony Fowler. 2021. *Thinking clearly with data: A guide to quantitative reasoning and analysis.* Princeton: Princeton University Press.

Kellstedt, Paul M., and Guy D. Whitten. 2018. *The fundamentals of political science research.* Cambridge: Cambridge University Press.