Week 5: Analysis of Proportions and Means

POP88162 Introduction to Quantitative Research Methods

Tom Paskhalis

Department of Political Science, Trinity College Dublin

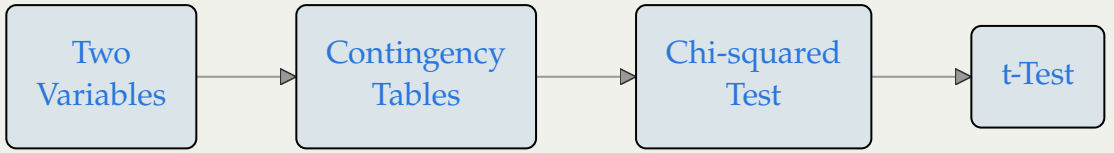
So Far

- *Sampling distributions* of population distributions of any shape approximate *normal distributions*.
- *Confidence intervals* describe how certain we are about the point estimates of population parameters.
- For *hypothesis testing* we formulate *null* and *alternative hypotheses*.
- We collect the data and calculate the test statistic to try to reject the null hypothesis.
- *P-value* based on calculated test statistic describes the weight of evidence against the null.

Topics for Today

- More hypothesis testing
- Joint probability distribution
- Conditional probability distribution
- Contingency table
- χ^2 test
- t-test

Today's Plan



Review: Hypothesis

- **Hypothesis** is a statement about the population.
- E.g. UK voters are equally likely to support Leave or Remain.
- We formulate two hypothesis:
  - H_0 - null hypothesis (typically, 0, indicating no difference/association)
  - H_a - alternative hypothesis (there is a difference/association)
- Alternative hypotheses can be *one-sided* or *two-sided*:
  - **one-sided**: the direction of difference/association is included
  - **two-sided**: no direction of difference/association is hypothesised

Review: Significance Test

- **Significance test** is used to summarize the evidence about a hypothesis.
- It does so by comparing the our estimates with those predicted by a null hypothesis.
- 5 components of a significance test:
  - **Assumptions**: scale of measurement, randomization, population distribution, sample size
  - **Hypotheses**: null H_0 and alternative H_a hypothesis
  - **Test statistic**: compares estimate to those under H_0
  - **P-value**: weight of evidence against H_0, smaller P indicate stronger evidence
  - **Conclusion**: decision to *reject* or *fail to reject* H_0.

Error Types

		Null hypothesis (H_0) is	
		True	**False**
Decision about Null hypothesis (H_0)	**Don't reject**	Correct inference (true negative) (probability = 1−α)	Type II Error (false negative) (probability = β)
	Reject	Type I Error (false positive) (probability = α)	Correct inference (true positive) (probability = 1−β)

When Do We Reject H_0?

- If the observed value of a test statistic is unlikely to occur by chance (given the null hypothesis)
- How low is very low?
- It depends on a chosen **confidence level**.
- Common levels are 0.95, 0.99 and 0.999.
- Alternatively, we can present it as **error probability** (α - alpha):

$$\text{Confidence Level} = 1 - \text{Error Probability} = 1 - \alpha$$

- With corresponding values of α: 0.05, 0.01 and 0.001.
- Remember, α-level indicates the accepted probability of committing Type I error.

Type I vs Type II Errors

- There is a trade-off between minimising Type I and Type II errors.
- As α-level (probability of Type I error) goes down, β-level (probability of Type II error) goes up.
- In other words, the harder it gets to reject H_0, the less likely we are to reject it even it is false.
- Choosing an α-level of 0.05 implies that we will falsely reject H_0 5% of the time under repeated sampling.
- Choosing an α-level of 0.01 implies that we will falsely reject H_0 1% of the time under repeated sampling.

Two Random Variables

Two Random Variables

- So far we discussed one random variable at a time.
- But most interesting questions in social sciences describe the relationships between two (or more) different variables.
- Recall, that our hypothesis would typically be:
  - Y is associated with X, or $X \rightarrow Y$
  - Y is associated with X conditional on Z, or $X \rightarrow Y : Z$
- In order to answer these question we need to understand the concept of **joint probability distribution** and **conditional probability distribution**.

Joint Probability Distribution

- For discrete variables, **joint probability distribution** describes the probability that two random variables (e.g. X and Y) simultaneously take some values (e.g. x and y).
- The probabilities of all possible combinations sum up to 1.
- Recall how we can write the probability of one random variable:

$$P(Y = y)$$

- For two random variables we can express their joint probability function as:

$$P(Y = y, X = x)$$

Conditional Probability Distribution

- For discrete variables, **conditional probability distribution** describes the probability of one variable (e.g. Y) taking different values, conditional on another random variable (e.g. X) having a specific value (e.g. x).
- The conditional probability of Y taking on some value y, when X takes values of x can then be expressed as:

$$P(Y = y|X = x)$$

- This probability is a joint probability of Y taking a value of y and X a value of x divided by the probability of X taking a value of x:

$$P(Y = y|X = x) = \frac{P(Y = y, X = x)}{P(X = x)}$$

Contingency Tables

Example: Brexit Referendum Poll

- Instead of looking at aggregate proportion of pro-Leave vs pro-Remain voters let's examine the distribution of those across 2 major parties: Conservative and Labour.
- We want to know whether there is an association between the party vote in 2015 General Election and support for leaving the EU in 2016 referendum.
- Or, in the language of probability theory — whether the conditional distribution of EU referendum vote is different depending on party support in 2015 General Election (GE).

$$P(Y_{EU\_2016} = y | X_{GE\_2015} = x)$$

- Or, plugging in the chosen values of x:
  - $P(Y_{EU\_2016} = y | X_{GE\_2015} = Conservative)$ is the probability of some EU referendum choice given the vote for Conservatives in 2015 GE.
  - $P(Y_{EU\_2016} = y | X_{GE\_2015} = Labour)$ is the probability of some EU referendum choice given the vote for Labour in 2015 GE.

Source

Fieldhouse et al. (2016), Hobolt (2016)

Contingency Table

- Two and more categorical variables tabulated together can be shown as a **contingency table** (or **crosstab** from **crosstab**ulation).
- The number of rows represents the number of levels for one categorical variable (X_{GE_2015}).
- The number of columns represents the number of levels for another categorical variable (Y_{EU_2016}).

		EU Referendum Vote		
		Leave	**Remain**	**Total**
Party Support in 2015 GE	**Conservative**	4289	2729	7018
	Labour	2112	4782	6894
	Total	6401	7511	13912

Joint Probability Distribution for Contingency Table

- Joint probability distribution describes relative frequencies of occurrences of different combinations of values:

$$P(Y_{EU\_2016} = y, X_{GE\_2015} = x)$$

		EU Referendum Vote		
		Leave	**Remain**	**Total**
Party Support in 2015 GE	**Conservative**	30.8% (4289)	19.6% (2729)	50.5% (7018)
	Labour	15.2% (2112)	34.4% (4782)	49.5% (6894)
	Total	46% (6401)	54% (7511)	100% (13912)

Conditional Probability Distribution for Contingency Table

- Conditional probability distribution describes sample data distribution of EU referendum vote *conditional on* vote in 2015 GE:

$$P(Y_{EU\_2016} = y | X_{GE\_2015} = x)$$

		EU Referendum Vote		
		Leave	**Remain**	**Total**
Party Support in 2015 GE	**Conservative**	61.1% (4289)	38.9% (2729)	100% (7018)
	Labour	30.6% (2112)	69.4% (4782)	100% (6894)
	Total	100% (6401)	100% (7511)	13912

Marginal Distributions

- In crosstabs row and column totals are called **marginal distributions**.

		EU Referendum Vote		
		Leave	**Remain**	**Total**
Party Support in 2015 GE	**Conservative**	30.8% (4289)	19.6% (2729)	50.5% (7018)
	Labour	15.2% (2112)	34.4% (4782)	49.5% (6894)
	Total	46% (6401)	54% (7511)	100% (13912)

Now Let's Set Up a Test!

- Null hypothesis:

  H_0: In the population vote choice in 2015 UK GE is independent from (has no association with) vote in 2016 UK EU membership referendum.

- Alternative hypothesis:

  H_a: In the population vote choice in 2015 UK GE is associated with vote in 2016 UK EU membership referendum.

- Let's ask ourselves, what would we expect to see in our data if our H_0 was true?
- And does our data look like that?

If Our Null Hypothesis Was True…

		EU Referendum Vote		
		Leave	**Remain**	**Total**
Party Support in 2015 GE	**Conservative**	46% (3229)	54% (3789)	100% (7018)
	Labour	46% (3172)	54% (3722)	100% (6894)
	Total	46% (6401)	54% (7511)	13912

- The cells of this contingency table show **expected frequencies** under H_0.
- Note that row percentages for each party are equivalent to the marginal distribution of votes in the EU referendum.

…but This Is What We See

		EU Referendum Vote		
		Leave	**Remain**	**Total**
Party Support in 2015 GE	**Conservative**	61% vs 46% (4289) (3229)	39% vs 54% (2729) (3789)	100% (7018)
	Labour	31% vs 46% (2112) (3172)	69% vs 54% (4782) (3722)	100% (6894)
	Total	46% (6401)	54% (7511)	13912

- The cells of this contingency table show **observed frequencies** as the first percentage (value).
- And the expected frequency (under the null hypothesis) as a second.

How Did We Calculate Expected Frequencies?

- Expected frequencies are a product of row and column totals for that cell, divided by the sample size.

$$f_e = \frac{total_{row} \times total_{column}}{n}$$

- For example to calculate the expected number of Conservative voters in 2015 GE choosing Leave under H_0:

$$f_e = \frac{7018 \times 6401}{13912} = 3229$$

Chi-squared Test

χ^2 Test Statistic

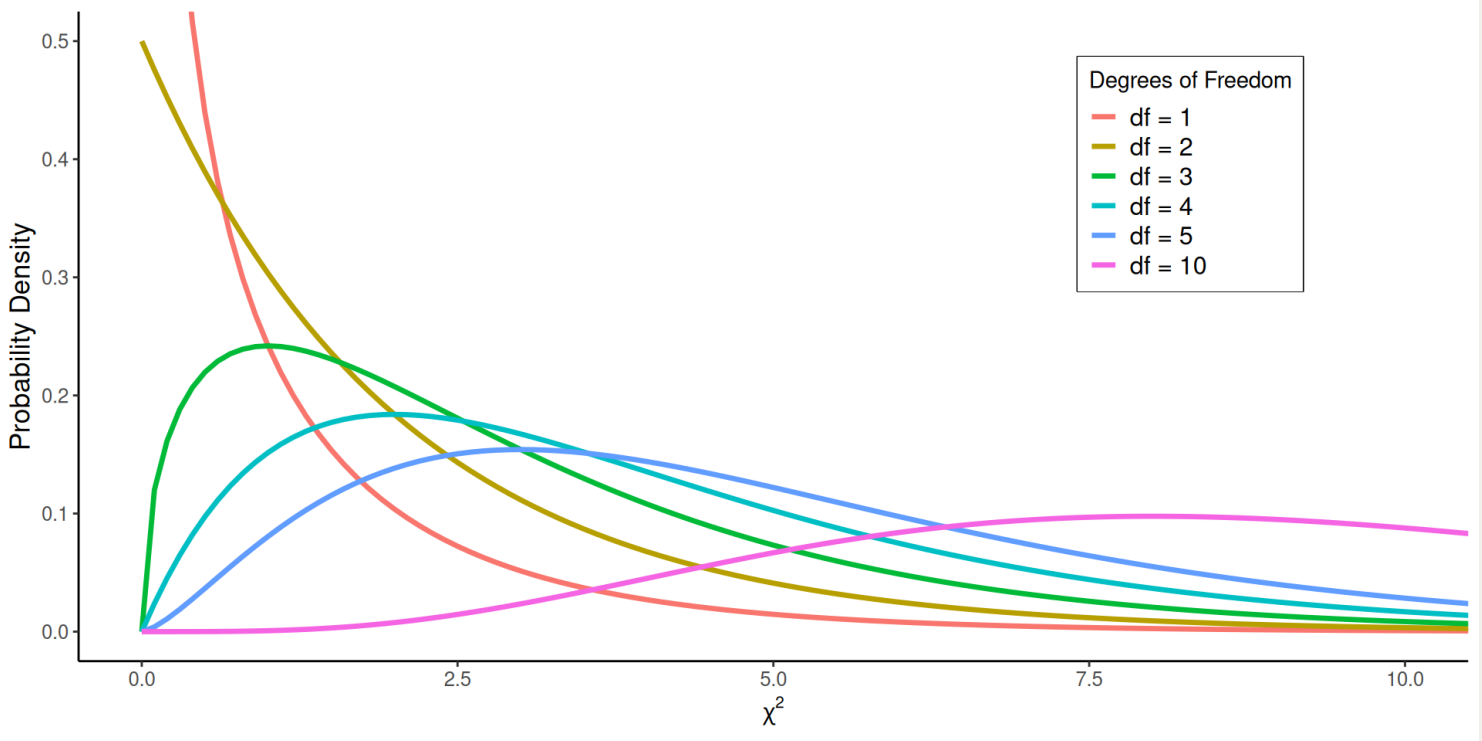
- We can use expected and observed frequencies to calculate a special statistic.
- Denoted by χ^2, it is called the **chi-squared statistic**.
- It summarises how close expected frequencies fall to observed frequencies:

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

- The summation is taken over cells in contingency table.
- The larger the value of χ^2, the greater the evidence against H_0: independence.
- This is the oldest test statistic in use today (introduced by Karl Pearson in 1900)

χ^2 Distribution

- χ^2 (chi-squared) distribution
- Recall how under Central Limit Theorem sampling distribution approximates normal.
- But for contingency table we only observe the cell frequencies.



Degrees of Freedom

- The only parameter describing the shape of a χ^2 distribution (technically, $\mu = df$ and $\sigma = \sqrt{2df}$).
- The higher the number of the degrees of freedom, the more the distribution is shift to the right, the larger the spread and the more it resembles a bell-shaped curve.
- How many cells can we fill as we like?

		EU Referendum Vote		
		Leave	**Remain**	**Total**
Party Support in 2015 GE	**Conservative**			7018
	Labour			6894
	Total	6401	7511	13912

Degrees of Freedom Continued

- The only parameter describing the shape of a χ^2 distribution (technically, $\mu = df$ and $\sigma = \sqrt{2df}$)
- The higher the number of the degrees of freedom, the more the distribution is shift to the right, the larger the spread and the more it resembles a bell-shaped curve.
- How many cells can we fill as we like?

		EU Referendum Vote		
		Leave	**Remain**	**Total**
Party Support in 2015 GE	**Conservative**	3229		7018
	Labour			6894
	Total	6401	7511	13912

- Well, one

Degrees of Freedom Continued

- The only parameter describing the shape of a χ^2 distribution (technically, $\mu = df$ and $\sigma = \sqrt{2df}$)
- The higher the number of the degrees of freedom, the more the distribution is shift to the right, the larger the spread and the more it resembles a bell-shaped curve.
- How many cells can we fill as we like?

		EU Referendum Vote		
		Leave	**Remain**	**Total**
Party Support in 2015 GE	**Conservative**	3229	3789	7018
	Labour			6894
	Total	6401	7511	13912

- Once we know the value in one, other must be filled with a particular value.

Degrees of Freedom Continued

- The only parameter describing the shape of a χ^2 distribution (technically, $\mu = df$ and $\sigma = \sqrt{2df}$)
- The higher the number of the degrees of freedom, the more the distribution is shift to the right, the larger the spread and the more it resembles a bell-shaped curve.
- How many cells can we fill as we like?

		EU Referendum Vote		
		Leave	**Remain**	**Total**
Party Support in 2015 GE	**Conservative**	3229	3789	7018
	Labour	3172		6894
	Total	6401	7511	13912

- Once we know the value in one, other must be filled with a particular value.

Degrees of Freedom Continued

- The only parameter describing the shape of a χ^2 distribution (technically, $\mu = df$ and $\sigma = \sqrt{2df}$)
- The higher the number of the degrees of freedom, the more the distribution is shift to the right, the larger the spread and the more it resembles a bell-shaped curve.
- How many cells can we fill as we like?

		EU Referendum Vote		
		Leave	**Remain**	**Total**
Party Support in 2015 GE	**Conservative**	3229	3789	7018
	Labour	3172	3722	6894
	Total	6401	7511	13912

- Once we know the value in one, other must be filled with a particular value.

Degrees of Freedom

- In general, when applying χ^2-test to a contingency table with r rows and c columns:

$$df = (r-1)(c-1)$$

- For example, for a 2×3 table, $r = 2$ and $c = 3$.
- Thus, $df = 1 \times 2 = 2$

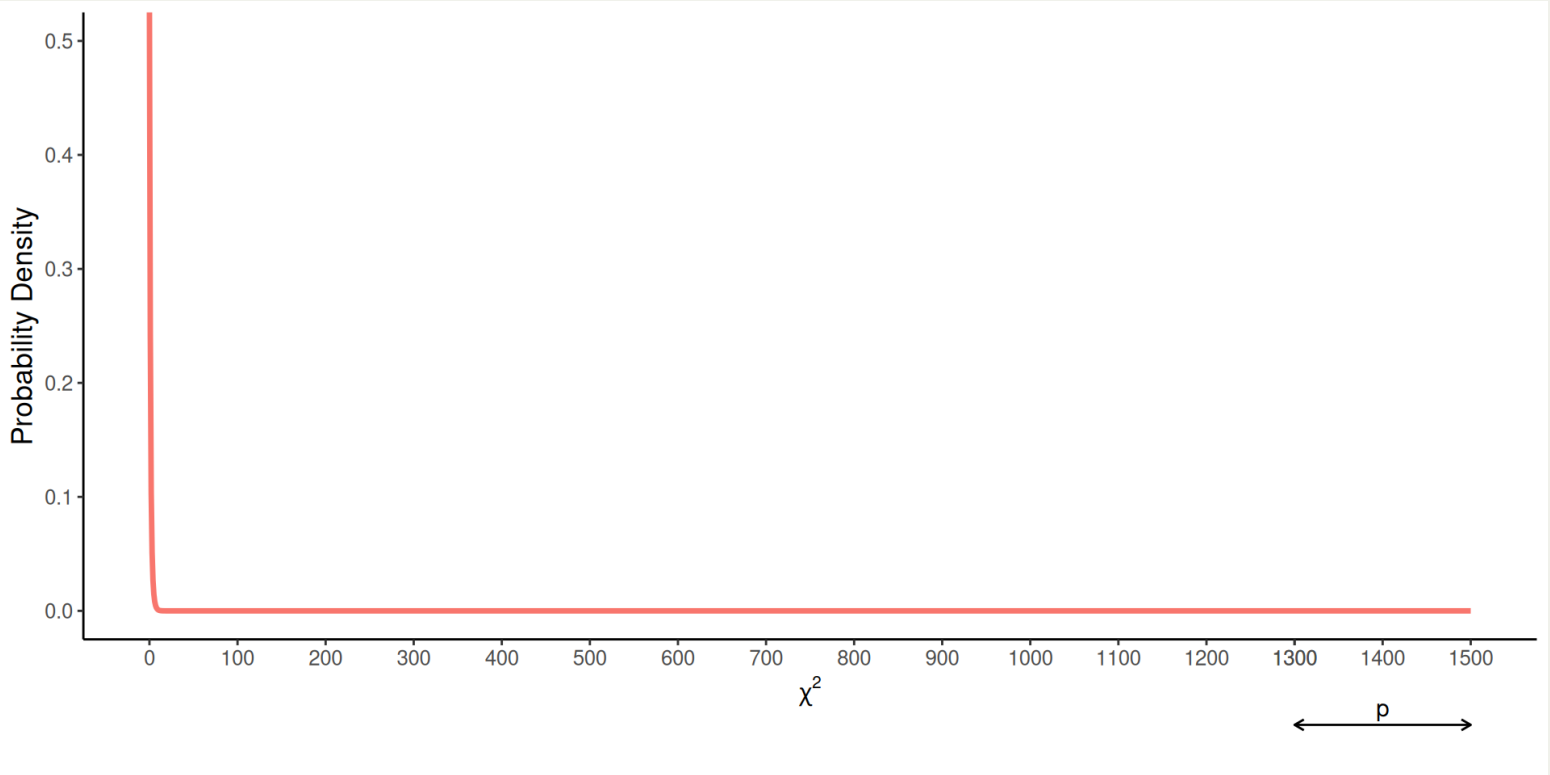
χ^2 Significance Testing

- Let's work out the χ^2 test statistic for the party vote in 2015 UK GE and preferences in 2016 EU membership referendum.

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e} = \frac{(4289 - 3229)^2}{3229} + \frac{(2729 - 3789)^2}{3789} + \frac{(2112 - 3172)^2}{3172} + \frac{(4782 - 3722)^2}{3722} \approx 1300$$

- And the number of the degrees of freedom is 1.

How Likely are We to Observe this Value?



- Very, very unlikely!

```
1 1 - pchisq(1300, df = 1)
```

```
[1] 0
```

What Conclusion Do We Make?

- With $p < 0.001$ we can reject the null hypothesis of no association between party vote in 2015 UK GE and vote choice in 2016 EU membership referendum at 0.1%-level.
- We can, thus, conclude that there is an association between these 2 variables in the population.
- In R we can carry out a χ^2 test by using `chisq.test()` function:

```
1 chisq.test(hobolt2016$vote_2015, hobolt2016$vote_eu)
```

```
    Pearson's Chi-squared test with Yates' continuity correction

data:  hobolt2016$vote_2015 and hobolt2016$vote_eu
X-squared = 1299.3, df = 1, p-value < 2.2e-16
```

Advantages of χ^2-test

- It's really simple — virtually no modelling assumptions at all.
- It works for any number of rows and/or columns, as long as the sample is relatively large.
- What does "relatively large" mean? At least 5 observations per cell. Otherwise, the sample statistic won't necessarily approximate a χ^2 distribution.

Disadvantages of χ^2-test

- Note that χ^2-test provides no indication about the strength of the association.
- An association can be statistically significant, but substantively inconsequential.
- Indeed, it may not even tell you the **direction** of the relationship, if that concept even makes sense (which it might not).

Statistical Tests

Statistical Tests

		Dependent Variable	
		Nominal/ Ordinal	**Interval**
Independent Variable	**Nominal/ Ordinal**	χ^2 (chi-squared) test	Mean comparison test
	Interval	Logistic Regression	Linear Regression

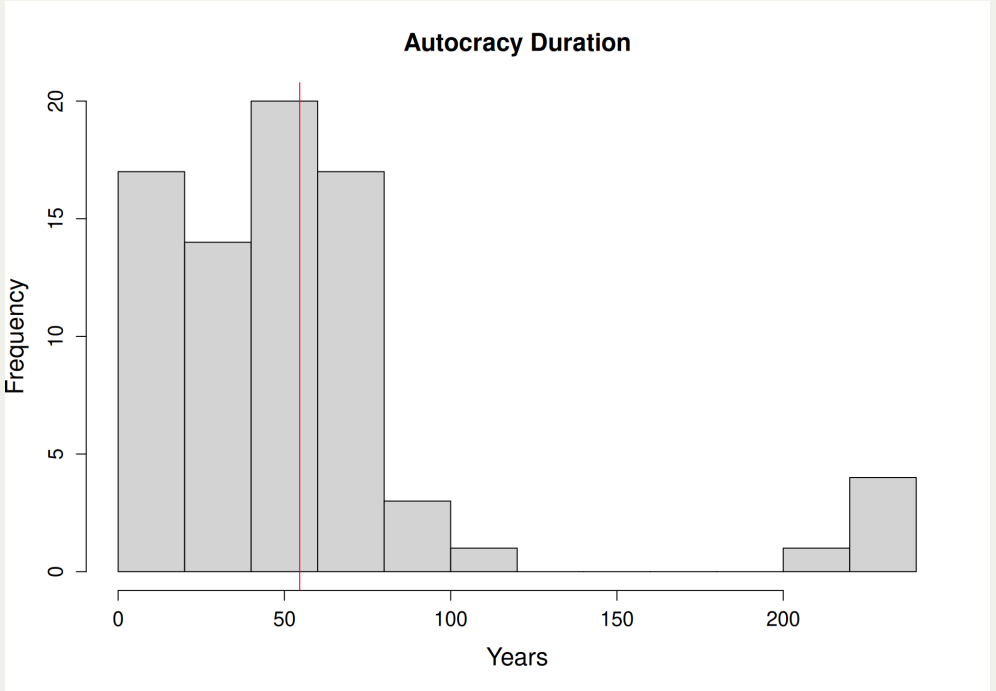
Difference in Means

- Oftentimes, we are interested in the difference between means in two groups.
  - Is women's income lower than men's income?
  - Do democracies last longer than autocracies?
- Let's denote one group as $Y_{X=0}$ and another group as $Y_{X=1}$.
- We are then interested in making inference about the difference:

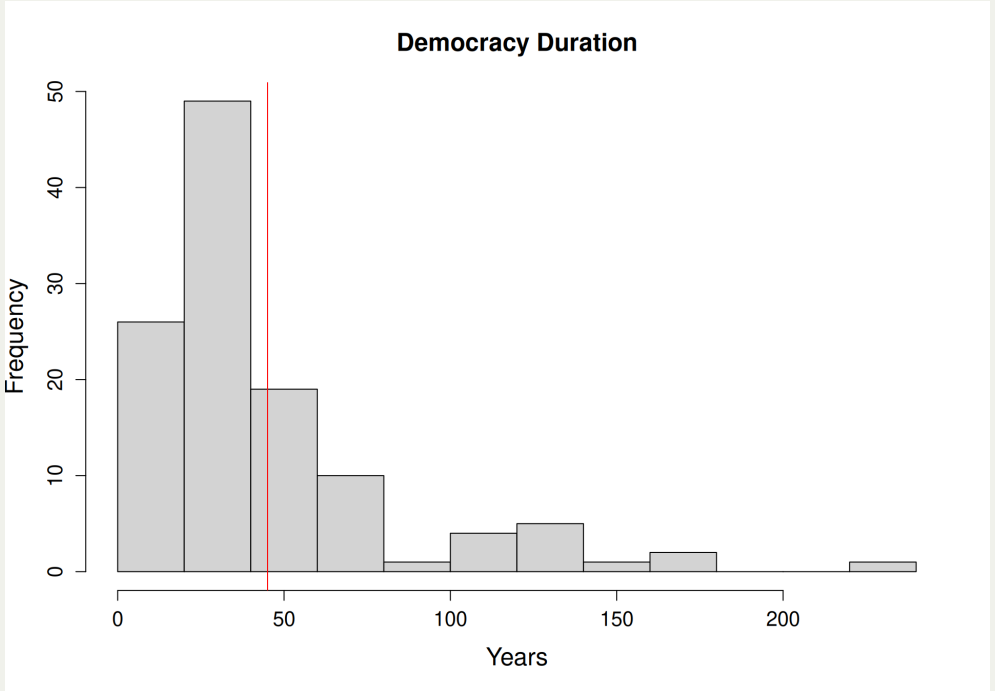
$$\bar{Y}_{X=0} - \bar{Y}_{X=1}$$

Example: Regime Longevity in 2020

Plot | Code



Plot | Code



Source

Boix, Miller and Rosato (2013), (2020)

Now Let's Set Up a Test

- Null hypothesis:

$$H_0 : \bar{Y}_{X=autocracy} = \bar{Y}_{X=democracy}$$

  or

$$H_0 : \bar{Y}_{X=autocracy} - \bar{Y}_{X=democracy} = 0$$

- Alternative hypothesis:

$$H_a : \bar{Y}_{X=autocracy} \neq \bar{Y}_{X=democracy}$$

- We will be testing out hypothesis at α-level of 0.05
- Recall the connection between confidence intervals and hypothesis testing.

Confidence Interval for Mean Difference

- For independent random samples from 2 groups that have normal probability distribution, a confidence interval is:

$$(\bar{Y}_{X=0} - \bar{Y}_{X=1}) \pm t(se)$$

- Where t is a t-score and standard error se:

$$se = \sqrt{\frac{s^2_{X=0}}{n_{X=0}} + \frac{s^2_{X=1}}{n_{X=1}}}$$

t-test Statistic

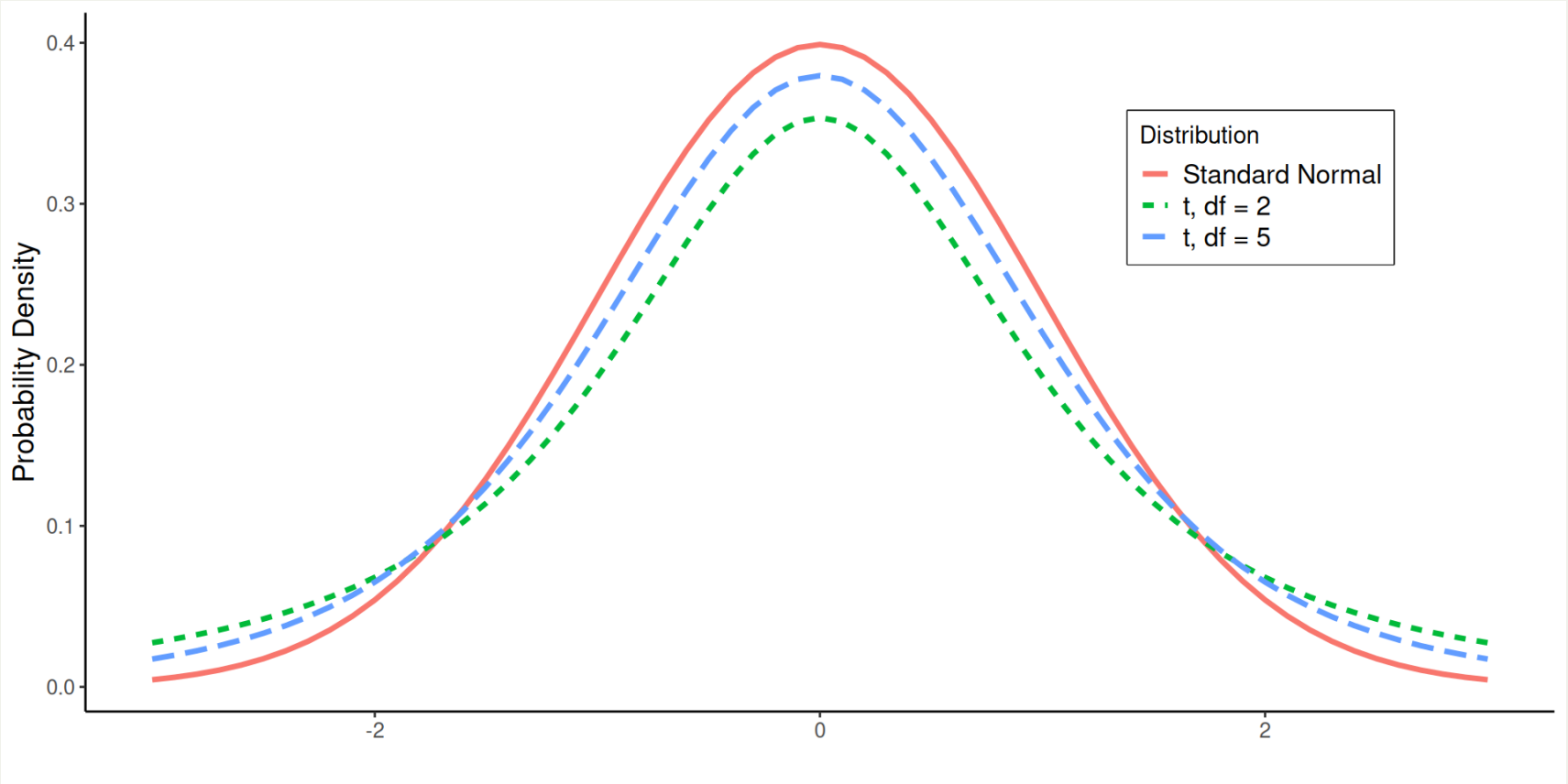
- Instead of calculating CIs, we can directly calculate a test statistic.
- Denoted by t, it is referred to as **t-test**.
- More specifically, for the difference in means (for a null hypothesis of no difference) it is:

$$t = \frac{\bar{Y}_{X=0} - \bar{Y}_{X=1}}{se_{\bar{Y}_{X=0} - \bar{Y}_{X=1}}} = \frac{\bar{Y}_{X=0} - \bar{Y}_{X=1}}{\sqrt{\frac{s^2_{X=0}}{n_{X=0}} + \frac{s^2_{X=1}}{n_{X=1}}}}$$

  - Where $s^2_{X=0}$ and $s^2_{X=1}$ are sample variances for 2 groups.
  - And $n_{X=0}$ and $n_{X=1}$ are the number of observations for each of the groups.

t Distribution

- t distribution is bell shaped and and symmetric around the mean of 0.
- In comparison to standard normal distribution the standard error is a bit larger than 1 and depends on the degrees of freedom.



Origins of Student-t Distribution

- Developed by William Gosset (under the pseudonym 'Student') while working at:

Degrees of Freedom

- The formula for the degrees of freedom for Student's t-distribution (often just called a t-distribution) is rather complex.
- Usually, it falls somewhere between $(n_{X=0} - 1) + (n_{X=1} - 1)$ and the minimum of $(n_{X=0} - 1)$ and $(n_{X=1} - 1)$.
- It is best left to R to correctly identify the degrees of freedom.

Example: Carrying Out t-test

- Let's work out the t-test statistic for the difference in longevity between autocracies and democracies.

$$t = \frac{\bar{Y}_{X=0} - \bar{Y}_{X=1}}{\sqrt{\frac{s^2_{X=0}}{n_{X=0}} + \frac{s^2_{X=1}}{n_{X=1}}}} \approx$$

$$\frac{54.61 - 45.05}{\sqrt{\frac{2498.004}{77} + \frac{1521.604}{118}}} =$$

$$\frac{9.56}{6.73} =$$

$$1.42$$

- Recall, $\sigma_{\bar{Y}} = \frac{\sigma}{\sqrt{n}}$ and $\hat{\sigma} = s^2 = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1}$
- For large sample sizes t-distribution approximates standard normal:

```
1 (1 - pnorm(1.42)) * 2
```

```
[1] 0.1556077
```

What Conclusion Do We Make?

- The probability of observing this difference under the null hypothesis is $\approx .158$
- Thus, we cannot reject the null hypothesis of no difference in longevity between autocracies and democracies at 5%-level.
- In R we can carry out a t-test by using `t.test()` function:

```r
t.test(democracy_2020$democracy_duration ~ democracy_2020$democracy)
```

```
    Welch Two Sample t-test

data:  democracy_2020$democracy_duration by democracy_2020$democracy
t = 1.4198, df = 134.61, p-value = 0.158
alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
95 percent confidence interval:
 -3.75709 22.87617
sample estimates:
mean in group 0 mean in group 1
       54.61039        45.05085
```

Next

- Workshop:
  - Factor Variables
- R Assignment 2 due:
  - **08:59 Tuesday, 25 February**
- Next week:
  - Correlation