- Final exam:
  - Tue Dec 14, P51 (together with course MVZb1003), probably 10am-11:40am
  - Wed Dec 15, 10am -11:24am

- The plan for today

- How do I know that the observed difference is big enough?
  - Learning about the population from the sample
    - Normal distribution
    - Central limit theorem
  - t-test

Science and uncertainty

- Scientific results – often there is a lot of uncertainty about them
  - Have we controlled for all possible confounding variables?
  - Do our measures represent our concepts well?
  - **Uncertainty arising from the fact that findings are based on a sample**

Learning about the population from the sample

- Population = ?
  - = all possible cases
- Sample = ?
  - = a subset of cases that are drawn from the underlying population
- Statistical inference = using what we know about the sample to infer (guess) what is likely to be true about the population
  - How? Using the properties of the **central limit theorem (CLT)**
    - Before learning about CLT, need to understand **normal distribution**

Normal distribution

- https://www.mathsisfun.com/data/standard-normal-distribution.html

- Bell-shaped curve
- Symmetrical (mean=median=mode)
- 68-95-99 rule

Are all variables normally distributed?

- No!
- But their *sample mean* is normally distributed!
  - Recall: sample mean v. population mean

- Roll a die 600 times and calculate the mean value

https://www.online-stopwatch.com/3d-dice/full-screen/

  - If it's a fair die, we know that the population mean 3.5
  - Our sample mean will be different – sometimes smaller, sometimes larger
  - Do this infinite number of times: "Roll a die 600 times"
  - **Sampling distribution** of the mean will be normally distributed

Image source: https://cheztacular.wordpress.com/2013/02/07/odds-of-rolling-a-dice/

Source: Kellstedt and Whitten (2009): The Fundamentals of Political Science Research. P. 124

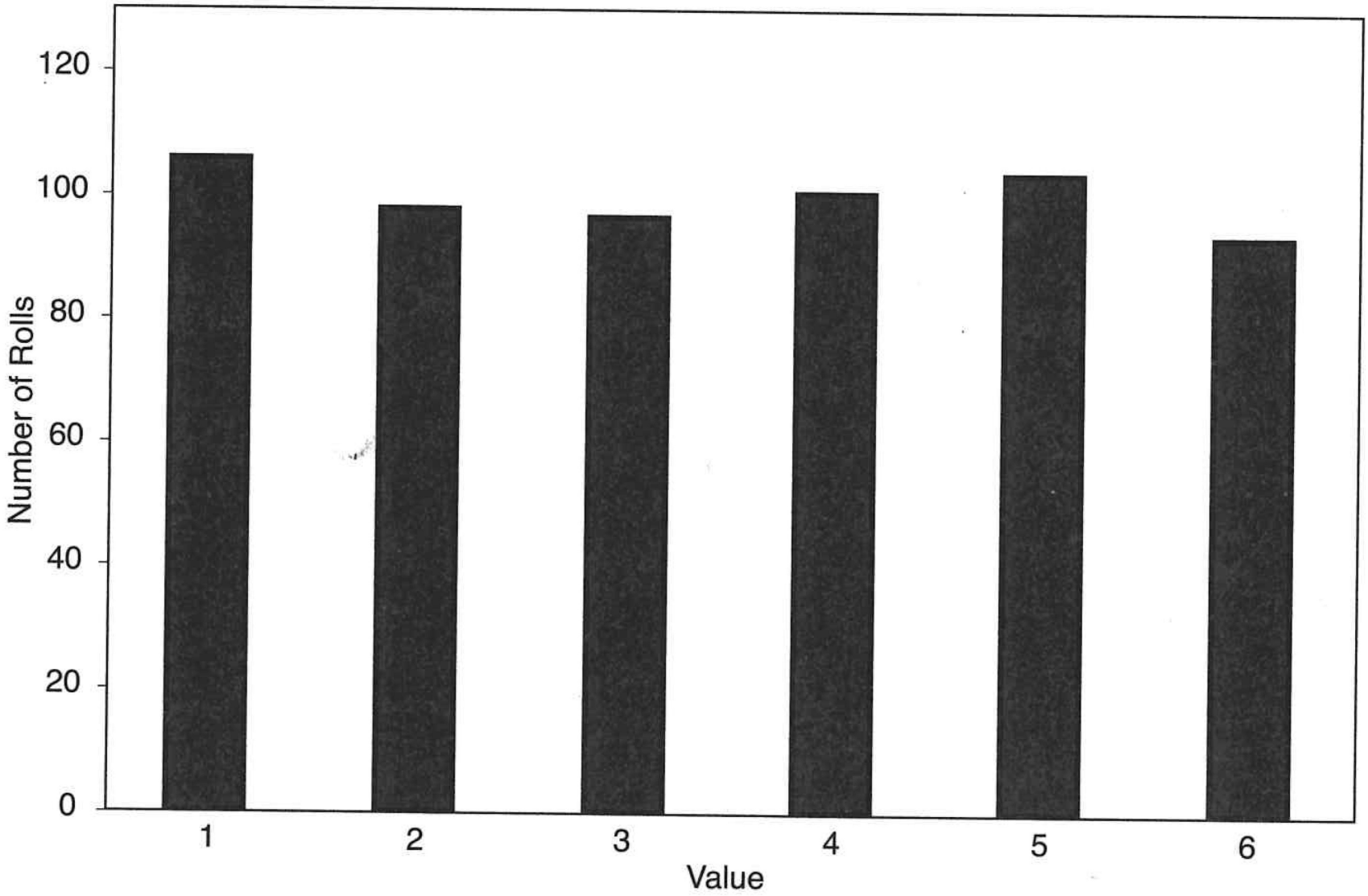


Figure 7.3. Frequency distribution of 600 rolls of a die.

Sampling distribution of the sample mean

- Normal distribution
- Centered on the true population mean
- Standard deviation of the sampling distribution
  - = standard error of the mean (or standard error)
  - = (std. deviation of the sample mean)/sqrt(sample size)
- Inference from the sample:
  - We are 95% confident that the true population mean is within ±(2*std. errors) from the sample mean
  - This range is the **confidence interval**

How to calculate the confidence interval for a sample mean in Jamovi?

- How many hours do Americans spend on the internet per week?
- GSS 2006 data – a sample. What is our estimate for the population?
- Go to *Analyses*--> *T-tests* --> *One Sample T-Test*. Put *wwwhr* in the *Dependent variable box.* Make sure that the following is selected in the options menu: *Student's*, *Mean difference* and *Confidence interval*, *Descriptives* (this one is optional) and that *0* is filled in the *Test value* box and that ≠ *Test value* is selected in the *Hypothesis options*. The treatment of missing values can be left at the default option (*Exclude cases analysis by analysis).*

- How many hours do Americans spend on the internet per week?
  GSS 2006 data – a sample. What is our estimate for the population?

One Sample T-Test

						95% Confidence Interval	
		Statistic	df	p	Mean difference	Lower	Upper
wwwhr	Student's t	29.4	2777	< .001	5.74	5.36	6.13

Descriptives

	N	Mean	Median	SD	SE
wwwhr	2778	5.74	1.00	10.3	0.196

Standard error

- "We are 95% confident that the true population mean lies between 5.36 and 6.13."
- 95% confidence interval = mean ± 2*std. error

Is the sample mean different from a certain number

- Hypothetical company – ask **all** employees – find that the average employee spends 5 hours a week on the internet
- Does the average American differ from the average employee in our company?
- Similar analysis as before, just put *5* into the *Test value* box.
- Interpretation:
  - A) Does the confidence interval include 0?
    - If not, the true mean is likely to be different from 5
  - B) Is the probability that we "got the *t statistic* of 3.81 just by chance" really small?
    - "really small" means "smaller than 0.05"
    - If the probability is smaller than 0.05, the mean is different from 5

One Sample T-Test

This is the probability value

One Sample T-Test

		Statistic	df	p	Mean difference	95% Confidence Interval	
						Lower	Upper
wwwhr	Student's t	3.81	2777	< .001	0.744	0.361	1.13

Note. H_a population mean ≠ 5

- Does the average American differ from the average employee in our company (who spend 5 hours a week on the web on average)?
- Interpretation:
  - The confidence interval is the CI for the difference between the sample mean and the *Test value* (5 in our case).
  - A) Does **the confidence interval** include 0?
    - If not, the true mean is likely to be different from 5
  - B) Is the probability that we "got the *t statistic* of 3.81 just by chance" really small?
    - "really small" means "smaller than 0.05"
    - If **the probability** is smaller than 0.05, the mean is different from 5

Do men and women *in the population* differ in their use of the internet?

- Variables: DV: wwwhr, IV: sex

- *Analyses --> T-tests --> Independent Samples T-Tests*
- Put *wwwhr* in the *Dependent variables* box. And *sex* in the grouping variable box.
- Make sure that the following is selected in the options menu: *Student's, Mean difference, Confidence interval 95%, Group 1 ≠ Group 2 (*in the *Hypothesis* section), *Homogeneity test (*in the *Assumptions Checks* section), and *Exclude cases analysis by analysis.*

Independent Samples T-Test

Independent Samples T-Test

							95% Confidence Interval	
		Statistic	df	p	Mean difference	SE difference	Lower	Upper
wwwhr	Student's t	3.94 [a]	2776	< .001	1.55	0.394	0.778	2.32

[a] Levene's test is significant (p < .05), suggesting a violation of the assumption of equal variances

Before any interpretation of the t-test, look at the test of the t-test's assumptions. If the p-value is smaller than 0.05, the assumption of equal variances in the two groups (men and women) is violated and Student's t-test cannot be used. You have to use Welch's test.

Assumptions

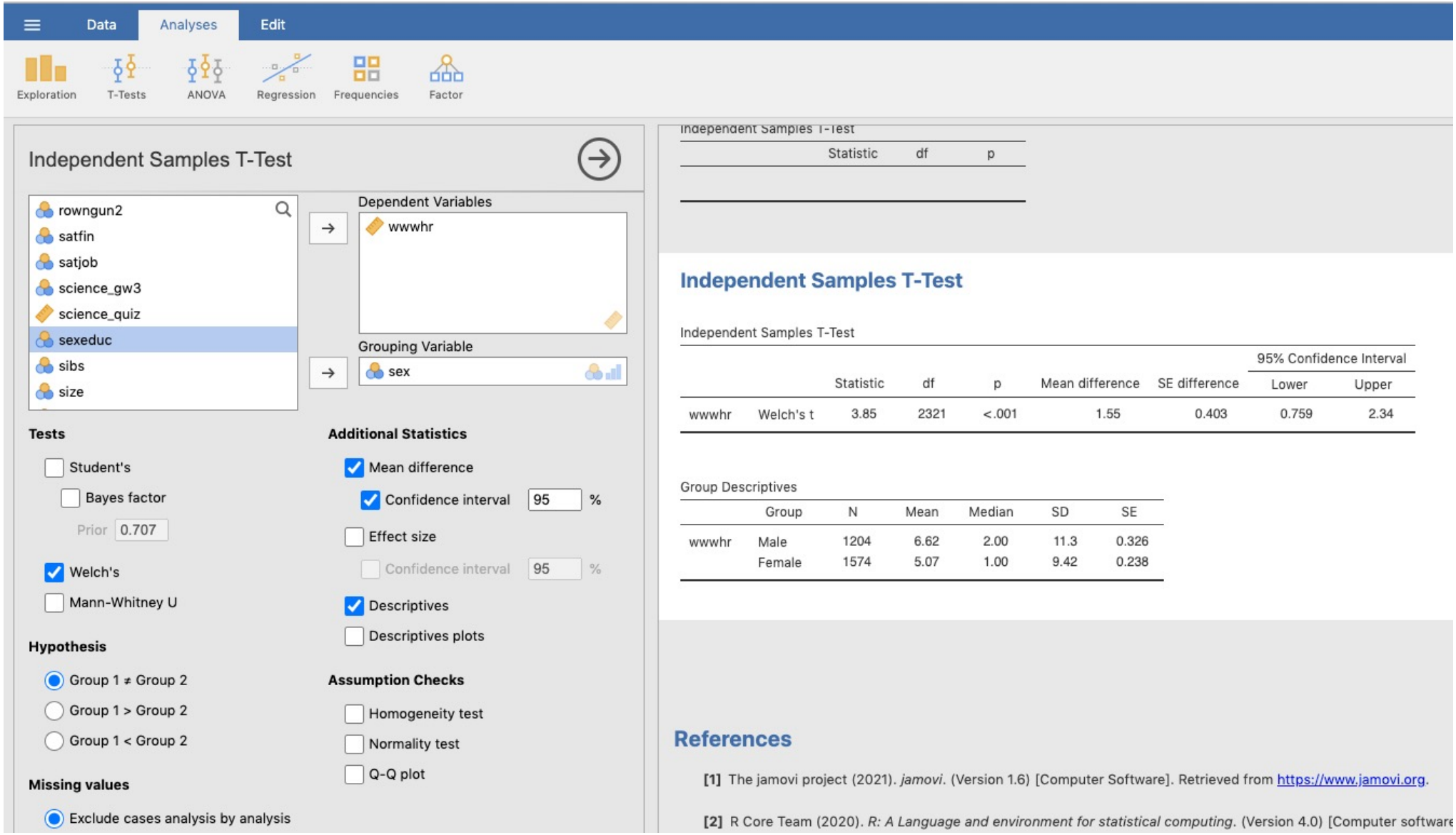
Homogeneity of Variances Test (Levene's)

	F	df	df2	p
wwwhr	22.5	1	2776	< .001

Note. A low p-value suggests a violation of the assumption of equal variances

- Our example: p-value is smaller than 0.05. We have to use Welch's test instead of the Student's t-test.
  - **Unselect Student's and select Welch's in the test's options.**
- If the probability were larger than 0.05, Student's t-test is fine and you can go ahead and interpret the Student's t-test.
- The interpretation of both of these tests is the same. The two t-tests differ a bit in how they are calculated but once you get the results, you interpret them the same way.

- How to set up and interpret the t-test: (next slides)



At this stage, you no longer need the Homogeneity test, you have already done the test.
Note we have selected the Welch's t-test instead of the Student's t-test.
We have asked for Mean difference, the 95% Confidence interval, and for Descriptives. The Descriptives are not required but usually it's a good idea to look at them as well.

Independent Samples T-Test

Independent Samples T-Test

		Statistic	df	p	Mean difference	SE difference	95% Confidence Interval	
							Lower	Upper
wwwhr	Welch's t	3.85	2321	< .001	1.55	0.403	0.759	2.34

Group Descriptives

	Group	N	Mean	Median	SD	SE
wwwhr	Male	1204	6.62	2.00	11.3	0.326
	Female	1574	5.07	1.00	9.42	0.238

Interpreting two-sample t-test

- Purple circle: the difference between the means (and its std. error to the right)
- A) Blue circle: the confidence interval (CI) for the difference between the means
  - If the CI includes zero, we cannot be confident that there really is any difference between the true population means. That is, we cannot be confident that men in American population spend more time on the internet than women do.

Interpreting two-sample t-test

- B) Brown circle: If this probability is smaller than 0.05, it is very unlikely that we would get this value of the *t statistic* if there really was no difference between the two population means
  - → we conclude that "We are 95% confident that there is a difference between the average number of hours that men and women spend on the internet"

- Once you establish whether there is a difference or not, you can proceed to comparing which group has higher/lower mean
  - Look at the numbers in the yellow circle and compare them
  - → Conclude that men spend more time on the internet

- A final note: Whenever you do a t-test in Jamovi, look at your dependent variable first in the Data tab. The measurement level on your dependent variable *must* be continuous. If it isn't, change the measurement level (Data→ Setup→ Measurement type--> Continuous)