Escritura académica en español 2: analizando datos

Fernanda Barrientos Contreras

18 de julio de 2018

(Materiales: ejercicios Excel.zip - bajar
 $\mathbf{aqu}\mathbf{i}$)

Ahora que ya sabemos utilizar Praat para obtener los datos que necesitamos, tenemos quwe analizar dichos datos. En este tutorial utilizaremos Excel para lograr los siguientes objetivos:

- 1. Obtener estadígrafos descriptivos sobre un determinado conjunto de datos.
- 2. Obtener estadísticas inferenciales de datos obtenidos.
 - Relación entre dos variables numéricas (continuas)
 - Comparación entre los promedios de dos grupos (variable continua vs. variable categórica)
 - Comparación entre los promedios de tres grupos (variable continua vs. variable categórica)
 - Comparación entre cuentas obtenidas por dos o más grupos (variable categórica vs. variable categórica)
- 3. Generar gráficos adecuados para el tipo de datos:
 - Gráfico de dispersión (scatterplot)
 - Carta de formantes
 - Gráfico de barras
 - Gráfico de cajas

1. Los estadígrafos descriptivos

Con el fin de explicar nuestros resultados, necesitamos describirlos y resumirlos. Consideraremos tres tipos esenciales:

- <u>Estadígrafos de concentración</u>: Muestran la tendencia en una muestra. La más conocida de estas medidas es la media aritmética (también conocida como promedio). De cualquier modo, el promedio por sí solo es un descriptor bastante pobre de los datos: no mostrará cuán dispersos están los datos, o la probabilidad de que un dato cualquiera tenga un valor X.
- Estadígrafos de posición: Son puntos de referencia contra el cual un punto específico puede ser comparado. La **mediana** es el descriptor más utilizado.

• <u>Estadígrafos de dispersión</u>: ¿Qué tan dispersos o concentrados están los datos en una muestra? ¿Qué tan lejos está un dato X del promedio? Esto puede cuantificarse calculando la **desviación estándar**.

1.1. Obtener estadígrafos descriptivos

Hay una manera muy simple de obtener todos los estadígrafos juntos. Para esto, abrimos vot.csv. En Excel 2016, vamos a *Daten > Datenanalyse > Populationskenngrößen* (Fig. 1).

Ē	-	o - C	~ .			1				vo	t - Excel							Ē	- 0
Da	atei	Start	Einfüger	n Seitenla	iyout Forme	eln Daten	Überprüf	en Ansicht	: 🗘 Wa									Anme	elden 🗸 Fre
Exter ab	ne Da rufen	ten N * Abf	leue rage + 🕞 Z	Abfragen an: Aus Tabelle Zuletzt verwe n und transfor	zeigen endete Quellen rmieren	Alle	Verb Eige Eige N - C. Verk	iindungen nschaften nüpfungen b dungen	earbeiten	A ↓ X A A↓ Sortieren	Filtern Trite	schen neut anwei weitert ern	nden Text in Spalter Dat	entools	Was-wäre-we Analyse * P	enn- Prognos rognose	eblatt Glieder	ung E	Datenanalyse 2 Analyse
F10)	-	: >	< 🗸 .	fx														
	,	A	В	С	D	E	F	G	н	1	J	к	L	м	N	0	Р	Q	R
1	subj	ect S	imple	Complex															
2	S1		27,71996	89,20183															
3	S1		28,80157	85,2692						1.4	ayi.			h					
4	S1		32,76409	100,6942			A	nalyse-Funkt	ionen				? ×						
5	S1		33,62823	72,52681				Analyse-Funkti	onen					1					
6	S1		21,36063	65,96122				Anova: Einfakt	orielle Varia	anzanalyse		^	OK						
7	S2		24,1324	82,28063				Anova: Zweifa	ktorielle Va	rianzanalyse mit	Messwiederholu	ng	Abbrechen						
8	S2		26,97996	87,02583				Anova: Zweita Korrelation	ktorielle Va	rianzanalyse ohn	e Messwiederhol	ung							
9	S2		37,43695	102,7169			3	Kovarianz					<u>H</u> ilfe						
10	S2		25,22999	85,65852			<u> </u>	Populationske	nngrößen										
11	S2		31,88534	59,42695				Zwei-Stichpro	ben F-Test										
12	S3		28,75428	118,2967				Fourieranalyse	e										
13	S3		37,12504	112,4412				Histogramm				~							
14	S3		29,74914	80,59898			_		-										

Figura 1: Calculando estadígrafos descriptivos.

Luego, una segunda ventana (Fig. 2) nos pedirá seleccionar el rango de datos (*Eingabebereich*) para lo cual hacemos clic en el recuadro (1): esto nos lleva de vuelta a la hoja de cálculo y ahí seleccionamos con el mouse todas las celdas. Luego seleccionamos dónde queremos tener la información: podemos crear una hoja nueva para eso (*Ausgabe > Neues Tabellenblatt*) (2). Le damos OK y en la hoja tendremos todos los datos deseados, junto a otros más.

Si no tenemos Excel 2016, tenemos las siguientes alternativas:

- Cargar el add-on Analysis Toolpak (Analyse-Funktionen). Esto es igual al botón Datenanalyse que aparece en Excel 2016.
- Tambien podemos calcular la media, la mediana y la desviación estándar haciendo clic en una celda vacía, luego clic en el ícono de funciones (ver Fig. 3a) y buscamos: MITTELWERT para promedio, MEDIAN para la mediana, y STABW.S para la desviación estándar (Fig. 3b). Luego hacemos clic en los recuadros pequeños (que nos llevan de vuelta a la hoja de datos) y seleccionamos los datos. El valor correspondiente aparecerá en la celda vacía donde hicimos clic al inicio.

2. Correlaciones entre dos variables continuas

En esta sección queremos saber si existe alguna relación entre la velocidad de lectura de una persona (medida en palabras por minuto o PPM) y el tiempo pasado frente a la TV (medido en horas por día). En el fondo, lo que queremos saber es si (a) existe una

6	ີ 5 ⁻	r → →	n Coitenla		ma ala Datan	Übereröfer	Anciela	O 14/20		vot - Excel	
xter	rne Daten	Abfrage - Da	n Seitenia Abfragen anz Aus Tabelle Zuletzt verwe en und transfor	zeigen endete Quel mieren	Ilen aktualisiere	Verbin Verbin Eigens N - Q Verkni Verbindu	idungen schaften üpfungen be	₩ was	A Z↓ Z↓ S	ortieren Filtern	Erne Erne Erwe
E3		>	< 🗸 .	fx							
	А	В	С	D	E	F	G	Н		I J	
1	subject	Simple	Complex	Pou	pulationskannar	ößen				2 ×	٦.
2	S1	27,71996	89,20183	FO	pulationskennigh	Jben				. ~	
3	S1	28,80157	85,2692	Ei	ngabe				-	ОК	
4	S1	32,76409	100,6942	Ēi	ingabebereich:	1					
5	S1	33,62823	72,52681	G	eordnet nach:		• Spalten			Abbrechen	
6	S1	21,36063	65,96122				O <u>Z</u> eilen			<u>H</u> ilfe	
7	S2	24,1324	82,28063	C	<u>B</u> eschriftungen i	n erster Zeile					
8	S2	26,97996	87,02583								
9	S2	37,43695	102,7169	Au	usgabe						
10	S2	25,22999	85,65852		Ausgabebereich						
11	S2	31,88534	59,42695		Neues Taballanh	1			_		
12	53	28,75428	118,2967		<u>neucs</u> rabellenic	2					
13	53	37,12504	112,4412		Neue Arbeitsma	ppe					_
14	53	29,74914	80,59898	[Statistische Kenr	ngrößen					
15	53	29,7811	97,87784		Konfidenzniveau	für Mittelwert		95	%		
10	53	34,32296	125,2355		k-größter Wert:		1				
17	54	30,47751	112 0221		k-kleinster Wert		1				-
10	54	20,00093	20 10920		mension were						-
20	54 CA	22,50310	15 21724					_	_		
20	54	20,00005	43,51724								
<u> </u>	54	33,05239	82,05240								

Figura 2: Estadígrafos descriptivos: selección de datos y del área para los resultados.

correlación; y (b) el sentido de la correlación (directa o inversa). En una **correlación directa**, mientras más <u>alto</u> sea el valor de una variable, más <u>alto</u> será el valor de la otra variable; en el caso de la **correlación inversa**, mientras más <u>alto</u> es el valor de una variable, más bajo será el valor de la otra variable.

2.1. Cálculo de R de Pearson

El test que nos sirve para este caso es la **R** de Pearson, que da valores entre -1 y 1. Si el número obtenido está entre -1 y 0, entonces la correlación es inversa; si el valor obtenido está entre 0 y 1, entonces la correlación es directa. Además, si el número se encuentra entre -1 y -0.5; o bien entre 0.5 y 1, entonces la correlación es alta.

Luego, también tenemos que reportar el valor p de la R de Pearson. En estadística, las pruebas arrojan una probabilidad de que lo que estamos observando es significativo. A esto se le llama valor p (p-Wert). En lingüística lo normal es asumir que un valor p menor a 0.05 es estadísticamente significativo, por lo que si obtenemos este valor entonces podemos decir en este caso que los resultados que obtuvimos son debido a una real correlación, con un 95 % de seguridad.

El archivo lectura.csv tiene una pequeña muestra con valores de velocidad lectura y horas frente a la TV. Notemos que en este caso un solo dato consiste en dos mediciones: persona A tiene una velocidad de lectura X y una cantidad de horas Y frente a la televisión.

Para encontrar la correlación con Excel, hacemos clic en una celda vacía, vamos al botón de funciones (Fig.3a), y buscamos PEARSON. Hacemos clic en los botones que nos llevan de vuelta a la hoja de cálculo (Fig. 4) y seleccionamos las celdas de la variable 1 (cantidad de horas frente a la TV). Volvemos al cuadro de diálogo pulsando el botón nuevamente, y luego seleccionamos las celdas de la variable 2 (velocidad de lectura). El valor aparecerá en la celda donde hicimos clic: -0.56. (Es convención redondear a 2

5	~ ¢~ •					- د• E		vo	ot - Exo
Datei	Start Einfüge	n Seitenlav	out Formel	n Daten	Überprüfe	Date: S	tart Einfügen Seitenlay	aigen Daten Überprüfen Ansicht V Was möchten Sie aigen all Verbindungen au Z	tun?
	_	,				Externe Daten	Neue Aus Tabelle	Alle Eigenschaften ZI Sortieren	Filter
		Abfragen anze	eigen		😒 Verbi	abruten *	Abfrage + Lo Zuletzt verwei	aktualisieren + Les verknuprungen bearbeiten	e automa
				0	The second se			versenangen	
erne Dat	en Neue	Aus labelle		Alle	Elger	E3	- × <)	x =	
la su chana a		7uletzt verwer	ndete Ouellen	L. P.	Verkr	A	вс	DEFGHI	
bruten *	Abtrage * LO	Lancie verver	acte Quenen	aktualisierer	Les rend	1 subject	Simple Complex	Funktion einfügen	?
	Abrufe	n und transforn	nieren		Verbind	2 51	27,71996 89,20183	Production and an	
						3 S1	28,80157 85,2692	Punkoon guchen:	
4		< 1 f	2			4 S1	32,76409 100,6942	Beschreiben Sie kurz, was Sie tun möchten, und klicken Sie dann auf 'OK'	ş
4		` ▼ J.	*			5 51	33,62823 72,52681	Kategorie auswählen: Statistik	
						7 52	24 1224 92 29062		
A	В	C	D	Ę	F	8 S2	26,97996 87,02583	Epikdon auswanien	
and a factor	et Cinenda	Conselan	Funktion einf	ügen		9 52	37,43695 102,7169	NORM.S.VERI	
subje	ct Simple	Complex				10 S2	25,22999 85,65852	PEARSON	
S1	27,71996	89.20183				11 S2	31,88534 59,42695	POISSON.VERT	
						12 S3	28,75428 118,2967	PROGNOSEETS	
51	28,80157	85,2692				13 53	37,12504 112,4412	PEAPSON/Matrix1:Matrix2)	
\$1	32 76409	100 6942				14 S3	29,74914 80,59898	Gibt den Pearsonschen Korrelationskoeffizienten zurück.	
191	52,70405	100,0042				15 53	29,7811 97,87784	-	
S1	33,62823	72,52681				17 54	30,47751 67,76986		
C1	21 26062	65 06122				18 54	25.66593 113.9321		
51	21,30003	05,50122				19 S4	22,56316 39,19829	Hilfe für diese Funktion OK	Abb
S2	24,1324	82,28063				20 S4	28,80005 45,31724		
62	26.07006	97 02592				21 \$4	33,85239 82,65246		
	20,97990	07,02003				22			

Figura 3: Calculando estadísticas descriptivas en versiones anteriores a Excel 2016.

decimales).

					_		vot -	Excel					_
en	Seitenlayout	Formel	n Daten	Überprüfen	Ansich	nt 🖓 Was							
Al A Zu fen	bfragen anzeige us Tabelle uletzt verwendet und transformiere	n :e Quellen en	Alle aktualisierer	Verbind Eigensch Verknüp Verbindung	ungen haften ofungen gen	bearbeiten	2↓ <u>Z A</u> Z↓ Sortieren I Sor	Filtern The Filtern	schen neut anwende veitert ern	n Text in Spalter Dat	entools	Was-wäre-w Analyse	renn- Progr
×	🗸 fx	=PEAR	5ON()										
	с	D	E	F	G	н	1	J	к	L	м	N	0
C	Complex												
6	89,20183												
7	85,2692												
9	100,6942	=P	EARSON()										_
3	72,52681					Funktionsarg	jumente					?	×
3	82,90122					PEARSON							
5	82,28003				-	Madelard							
5	102.7169					Maurixi				- array)		
9	85.65852					Matrix2				= array			
4	59,42695									-			
8	118,2967					Gibt den Pears	onschen Korrelati	onskoeffiziente	en zurück.				
4	112,4412							Matrix1	ist eine Reih	e von unabl	längigen Wer	ten.	
4	80,59898												
1	97,87784												
6	125,2355					Formelergebni	is .						
1	67,76986					. errergebri							
3	113,9321					Hilfe für diese	Funktion				ОК	Abbr	echen
6	39,19829				L	_							
5	45,31/24												
9	82,05246												

Figura 4: Calculando R de Pearson.

Pero también necesitaremos reportar el valor p en el informe. Para eso, tenemos que hacer algo un poco prehistórico (Excel no es lo mejor para análisis de estadísticas!): chequear el valor p en una tabla de valores críticos de R de Pearson (ver tabla adjunta).

Cómo leemos la tabla? Primero buscamos el número de pares que analizamos (en este caso, 15). Miramos en la fila correspondiente, en la columna .05. Si nuestro valor de R es mayor al que aparece ahí, entonces bingo! Nuestra correlación es significativa a un nivel de seguridad de 95 %.

Finalmente, el reporte se escribe así:

Se calculó un coeficiente R de Pearson para determinar si existe una correlación entre la velocidad de lectura y las horas frente a la TV. La prueba arrojó significados significativos: r (13) = -0.56 (p < 0.05). La cantidad de horas frente a la TV se encuentra inversamente relacionada con la velocidad de lectura.

IMPORTANTE:

- El reporte de un test estadístico DEBE TENER SIEMPRE:
 - El valor del test estadístico
 - Los grados de libertad (van entre paréntesis)
 - El valor p (se indica si es menor a 0.05, 0.01, o 0.001)
- Los grados de libertad para la prueba R de Pearson son simplemente el número de pares analizados -2.

2.2. Gráfico de la correlación

Para generar el gráfico de la correlación, vamos a necesitar un gráfico XY o scatterplot. Primero, seleccionamos todos los datos. Luego vamos a Einfügen > Empholene Diagramme > Punkt (XY) (Fig. 4). En otras versiones de Excel podemos ir a la pestaña Einfügen > Punkt.



Figura 5: Creando un gráfico de tipo XY.

Pero esto nos dará solamente los puntos en el gráfico, y no la línea de correlación. Para eso vamos a *Entwurf > Diagrammelement hinzufügen > Trendlinie > Linear* (Fig. 6). En versiones anteriores de Excel, vamos a *Layout > Trendlinie > Lineare Trendlinie*. Luego podemos (jy debemos!) personalizar nuestro gráfico haciendo clic en él y usando las opciones de la pestaña *Entwurf*.

3. Comparación entre dos grupos (variable numérica vs. variable categórica)

Ahora suponemos que queremos comparar la duración del VOT en ataques (onsets) silábicos simples (esto es, solamente una consonante al comienzo de la sílaba) vs. en

6	5 -6	⇒ =			reading - Ex	cel			Diagra	mmtools						
Da	itei Stari	t Einfüger	n Seitenla	ayout For	meln Dat	en Über	prüfen	Ansicht	Entwurf	Format	ਊ Wa					
Diag hi	rammelemen nzufügen •	nt Schnelllay	out Farber änderr	n n					Dagumentai	5xx		Depended 		Depended 		•
db	Ach <u>s</u> en		•						Diagrammt	ormatvorlage	en					
dib.	Achsentitel		· 🗸	fx												
dis dis dis	Diagra <u>m</u> mti Daten <u>b</u> esch Datentabelle	tel riftungen e	C PM .0,5022	D	E	F	G		н	I	J	К	L	М	N	
曲期	F <u>e</u> hlerindika <u>G</u> itternetzlin	toren ien	, 57,6686 , 55,7731)6,0719				o									
db Ø	Lege <u>n</u> de Linien		2,1885 7,8462				300			Di	agram	mtitel				Cred -
<u>∠</u> ⊡	<u>Trendlinie</u> Pos./Neg. A	bweichung	, 🟒	<u>K</u> eine			250				•	••	•			Y
11	10	2,069054	1	Linear			200				•	• • • • • • •				
12	11	1,59911	2:	Exponentia	I		150				•					
14 15	13	1,459804 1,921233	18	Lineare <u>P</u> ro	gnose		100									
16 17	15	1,46412	19	Gleitender	Durchschnitt	t	50									
18 19			We	eitere Trendli	nienoptione	n	0	0	0,5	1	0	,5	2	2,5	3	
20																

Figura 6: Línea de correlación.

ataques complejos (dos o más consonantes al comienzo de la sílaba) en un idioma X. El archivo VOT.csv tiene una pequeña muestra.

3.1. Cálculo de t-test

Entonces queremos comparar: ¿hay una diferencia significativa entre el VOT promedio de los ataques simples vs. el VOT de los ataques complejos? Para esto necesitamos un **t-test**, o T de Student. Este test compara los promedios y nos da un valor p.

Hay distintos tipos de t-test:

- Pareado o no pareado (Ein Stichprobe (= Gepaart) / Zwei Stichproben): Se hizo la medición en el mismo grupo de personas o en dos grupos distintos? Si medimos algo en el mismo grupo de personas (como en este caso), entonces hacemos un t-test pareado. Por ejemplo, si los datos en VOT.csv fueran de 8 sujetos, donde 4 corresponden a la condición Ataque Simple y 4 a la condición Ataque Complejo, entonces necesitamos un t-test no pareado.
- De una o dos colas (einseitiger / zweiseitiger): ¿Tenemos una hipótesis sobre la dirección en la cual se da la diferencia? (i.e. si la condición Ataque Simple tiene un promedio MENOR al de la condición Ataque Complejo)? Si tenemos una idea sólida, entonces podemos correr un t-test de una cola. Si no, la mejor opción es un test de dos colas (y esto es lo que siempre se sugiere).

Para calcular el t-test, tenemos dos alternativas:

 Podemos ir a Daten > Datenanalyse y luego seleccionar el t.test (ver item 2) correspondiente. Seleccionamos los datos y tenemos la tabla con todos los valores necesarios en un solo paso. Para reportar necesitamos el valor de la prueba misma (Fig. 7, celda amarilla), los grados de libertad (celda naranja) y el valor p (celda verde). • Hacemos clic en una celda vacía (por ejemplo, E5), vamos a ir al botón de funciones (Fig. 3a), y buscamos T.TEST. Aquí tenemos que llenar los campos como lo hicimos con la R de Pearson, pero además tenemos que establecer si queremos un test de 1 o 2 colas (einseitiger / zweiseitiger), y luego si es pareado (gepaart), no pareado homocedástico (zwei Stichproben, gleiche Varianz - homoskedastisch), o no pareado heterocedástico (zwei Stichproben, ungleiche Varianz - heteroskedastisch) (Fig. 7). En este caso, haremos un t-test con dos colas (Escribimos 2 en el campo *Seiten*) y pareado (Escribimos 1 en el campo *Typ*). En la celda aparecerá solamente el valor p: 4.88E-10. ($4.88^{-10} = 0.00000004.88 < 0.05$, o sea, tenemos un resultado significativo, jyay!).

∎∽∂	- -					vot - Excel					8	- 6	×
Datei Start	Einfügen S	eitenlayout f	Formeln Da	ten Überprüf	en Ansicht 🖓 Was mö							elden 🔉 Fi	eigeben
Einfügen 💉	Calibri • FKU•	11 • A A • <u>A</u> • <u>A</u> •		≫ - PText Text ■ E Vert Ausrichtur	tumbruch binden und zentrieren v	r % 000 00 000 Beding Zahl 5	te Als Tabelle Ze ung * formatieren * Formatvorlager	llenformatvorlagen	Einfügen Köschen Format Zellen	×Σ× •↓ • •	A Sortieren und Filtern * Bearbeit	Suchen und Auswählen *	~
E6 ~	: ×	√ fx											~
A	в	C	D	F	F	6	н	1		к	1.1	м	
1 subject	Simple	Complex											
2 S1	27,7199594	89,2018275											
3 S1	28,8015703	85,2692017											
4 S1	32,7640908	100,694233											
5 S1	33,6282282	72,5268101											
6 S1	21,3606279	65,9612225			1								
7 S2	24,1323965	82,2806317			Zweistichproben t-Test b	ei abhängigen Stichprobe	n (Paarvergleichste	st)					
8 S2	26,9799563	87,0258262											
9 S2	37,436948	102,716902				Variable 1	Variable 2						
10 S2	25,2299899	85,6585223			Mittelwert	29,55153335	85,70411497						
11 S2	31,8853404	59,4269531			Varianz	20,25983953	538,4076798						
12 S3	28,7542819	118,29672			Beobachtungen	20	20						
13 S3	37,1250404	112,441155			Pearson Korrelation	0,41377832							
14 S3	29,7491396	80,5989786			Hypothetische Differenz	de 0							
15 S3	29,7811022	97,8778355			Freiheitsgrade (df)	19							
16 S3	34,3229595	125,235527			t-Statistik	-11,55593544							
17 S4	30,4775135	67,7698579			P(T<=t) einseitig	2,44249E-10							
18 S4	25,6659252	113,932106			Kritischer t-Wert bei ein:	seit 1,729132812							
19 \$4	22,5631599	39,1982898			P(T<=t) zweiseitig	4,88497E-10							
20 \$4	28,8000467	45,3172416			Kritischer t-Wert bei zwe	eise 2,093024054							
21 \$4	33,8523906	82,652459											
22													-
< > VC	ot 🕀						÷ •						Þ
Bereit											─	-1	+ 100 %
t di	🤌 📄			0						я ^я ^	🖿 🧟 🗄	16:24 17.07.2018	5

Figura 7: Calculando un t-test.

Ahora, necesitamos el valor de t propiamente tal. Hacemos clic en una celda vacía y escribimos lo siguiente:

... Que da un valor de 11.55. En esta función usamos el valor p (que estaba en la celda E5) y los grados de libertad, que en este caso son 19 (número de observaciones -1). Entonces reportamos de la siguiente manera:

Se realizó una prueba t
 de Student pareada de dos colas para determinar si la diferencia entre los valores de VOT en Ataque Simple es diferente a la de la condición Ataque Complejo. Los resultados son significativos:
 t~(19) = 11.55~(p > .05). Esto sugiere que en el idioma X, los valores de VOT en ataques simples son significativamente más largos que los de los ataques complejos.

3.2. Gráfico de variables numérica vs. variable categórica

Aquí tenemos dos posibilidades. Podemos hacer un gráfico de cajas (*Kastengrafikdia-gramm* o *Boxplot*) o un gráfico de barras (*Säulendiagramm*). Lamentablemente, sólo Excel 2016 tiene la posibilidad de hacer un gráfico de cajas; la otra posibilidad es hacer un gráfico de barras y disfrazarlo de gráfico de cajas (¡muy complicado!). Por razones de tiempo,

vamos simplemente a aprender a hacer un gráfico de barras (pero en Excel 2016 basta con seleccionar los datos, y luego ir a $Einfügen > Empholene \ Diagramme > Kastengrafik$ (Fig. 8) para tener un diagrama de cajas).



Figura 8: Diagrama de cajas (sólo en Excel 2016)

Para hacer un gráfico de barras, seleccionamos los datos de vot.csv, y luego vamos a *Einfügen* > *Empholene Diagramme* > *Säulendiagramm*. El diagrama aparecerá automáticamente. Luego tenemos que poner las barras de error: estas dan una idea de qué tan representativa *de la población entera* es el promedio de la muestra. Las barras de error son MUY importantes en los gráficos científicos. Para esto necesitamos calcular el margen de error al 5 % (nuevamente, convención). Esto se hace de la siguiente manera:

ۍ 🖬	¢				- د ب ا	¢- ∓			
Datei Si	art Einfügen Seitenla	ayout Form	ıeln Daten Üb	erpri D	atei Sta	art Einfüger	n Seitenlay	out For	meln
1 × ×	Calibri 7 11	· ^ ~ =	= &.	De Te	🚔 👗				
						Calibri	* 11 *	AA	_
Lintugen	F K U - ⊞ - ≤	<mark>≥ - A</mark> - ≣		Ein Ein	fügen 🚆	FKU	• 🗄 • 🕭	- <u>A</u> -	83
vischenablage	S Schriftart	6	Aus	richt	· · · ·			_	
		£ (20)		ZWIS	chenablage	* >	chrittart	19	
824		Jx =(B23	3/WURZEL(20))	B2	5	- : ×	 ✓ j. 	x =1,9	6*B
A	B C	D	E	F	d .	-	-	_	
6 S1	21,36063 65,96122				A	B	0.000	U	
7 S2	24,1324 82,28063			6	51	21,36063	65,96122		
8 S2	26,97996 87,02583			7	52	24,1324	82,28063		
9 S2	37,43695 102,7169			8	S2	26,97996	87,02583		
10 S2	25,22999 85,65852			9	S2	37,43695	102,7169		
I1 S2	31,88534 59,42695			10	S2	25,22999	85,65852		
2 \$3	28,75428 118,2967			11	S2	31,88534	59,42695		
I3 S3	37,12504 112,4412			12	S3	28,75428	118,2967		
14 S3	29,74914 80,59898			13	S3	37,12504	112,4412		
5 \$3	29,7811 97,87784			14	S3	29,74914	80,59898		
6 S3	34,32296 125,2355			15	S 3	29,7811	97,87784		
7 S4	30,47751 67,76986			16	S 3	34,32296	125,2355		
8 S4	25,66593 113,9321			17	<u>\$4</u>	30 47751	67 76986		
9 S4	22,56316 39,19829			18	<u>\$4</u>	25.66593	113,9321		
20 S4	28,80005 45,31724			19	54	22 56316	39 19829		
1 S4	33,85239 82,65246			20	54	28 80005	45 31724		
22 prom	29,55153 85,70411			20	54	22,30000	97 65746		
23 stdev	4,501093 23,20361			21	34	35,85239	02,03240		
4 sem	1,006475 5,188486			22	prom	29,55153	85,70411		
5 margin	1,972691 10,16943			23	stdev	4,501093	23,20361		
26	1,972691 10,16943			24	sem	1,006475	5,188486		
27				25	margin	1,972691	10,16943		
	vot vot2 Sheet3	(Ŧ)		26		1,972691	10,16943		

(a) Cálculo del error estándar

(b) Margen de error

Figura 9: Calculando error estándar y margen de error. La fórmula aparece en la barra de funciones.

1. Calculamos el error estándar (Fig. 9a): Supongamos que tenemos la desviación estándar de la variable Simple en la celda E20. En una celda vacía (supongamos, E21) escribimos:

=E20/WURZEL(20)

Es decir, la desviación estándar dividida por la raíz cuadrada del número de observaciones. Repetimos con los valores de la variable Complejo.

2. Calculamos el margen de error (Fig. 9b): En una celda vacía (por ejemplo, E22) escribimos

=E21*1,96

Repetimos con la otra variable.

Ahora volvemos al diagrama. Hacemos clic en el y vamos a *Diagrammelement hinzufugen* > *Fehlenindikatoren* > *Weitere Fehlerindikatorenoptionen....* Seleccionamos *Beide, Mit Abschluss,* y luego *Benutzerdefiniert.* Hacemos clic y en *Positiver Fehlerwert* hacemos clic al botón que lleva a la hoja de cálculo y seleccionamos donde tenemos los márgenes de error. Luego, en *Negativer Fehlerwert* hacemos exactamente lo mismo (mismas celdas) (Fig. 10). Pulsamos OK y el diagrama tendrá barras de error.

									Diagram					
D	atei Start	t Einfügen	Seitenlayout	Formeln	Daten	Überprü	fen	Ansicht	Entwurf	Format	🖓 Was möchte	en Sie tun		
Dia <u>c</u> h	grammelemer hinzufügen * Diagramm	nt Schnelllayo	out Farben ändern	Typester	25533.3300 0 0	errita 	0 145 E 4	AMETTI.	Diagrammfor	rmatvorlager			Disparent	
B2	5	• : ×	$\checkmark f_x$											
	A	В	C [2	E	F		G	н	1	J	к	L	
6	S1	21,36063	65,96122			C	,			5.				
7	S2	24,1324	82,28063							Dia	igrammtitel			
8	S2	26,97996	87,02583				120							
9	S2	37,43695	102,7169											
10	S2	25,22999	85,65852				100						φ	
11	S2	31,88534	59,42695			↓	80							
12	53	28,/5428	112,4412			+							••	
13	33 52	37,12504	112,4412				60							
14	33	29,74914	07,29098						Annual Contract	a ratio e e				
16	53	34 32296	125 2355				40	Benu	uzeraetiniert	te Fenlerind	I ? X	`		
17	S4	30,47751	67.76986				20	Positi	ver Fehlerwer	t				
18	S4	25,66593	113,9321					=vot	!\$B\$25:\$C\$	E.				
19	S4	22,56316	39,19829				0	<u>N</u> ega	tiver Fehlerwe	ert				
20	S4	28,80005	45,31724					=vot	!\$B\$25:\$C1	E.			2	
21	S4	33,85239	82,65246			C	<u> </u>			ОК	Abbrechen			
22	prom	29,55153	85,70411					L						
23	stdev	4,501093	23,20361											
24	sem	1,006475	5,188486											
25	margin	1,972691	10,16943											

Figura 10: Llenando el campo de valores para las barras de error.

4. Comparación entre 3+ grupos (variable numérica vs. variable categórica)

Ahora vamos a suponer que tenemos los datos no solamente de la duración del VOT en ataques simples y complejos, sino también en coda (final de sílaba). En el archivo vot2.csv hay una pequeña muestra. Qué tan diferente entonces son los VOT en estos tres grupos?

4.1. Cálculo de ANOVA

Para medir esto necesitamos hacer un análisis de varianza (ANOVA). Si tenemos Excel 2016 o el add-on Analysis Toolpak, entonces podemos ir a Data > Datenanalyse y buscar la función ANOVA (*Varianzanalyse*). Pero hay varias posibilidades:

- Unifactorial (einfaktorielle) o bifactorial (zweifaktorielle): Queremos medir el efecto de UNA SOLA VARIABLE o DOS VARIABLES sobre 3+ grupos? Si es una sola variable, entonces usamos ANOVA unifactorial.
- Con muestras repetidas (mit Messwiederholung): ¿Queremos medir efectos sobre UN MISMO GRUPO que ha sido medido 3+ veces (como el t-test pareado con una muestra, pero con 3+ mediciones sobre los mismos sujetos)? Si es así, entonces necesitamos una prueba con muestras repetidas.

Una vez que ya sabemos qué test realizar (en el caso de vot2.csv sería unifactorial) Entonces simplemente buscamos la función y luego ingresamos los datos, pulsando el botón que lleva a la hoja de cálculo y luego seleccionando todos los datos. Esta función no está disponible en el botón de funciones; solo es parte del botón *Datenanalyse*.

Aquí necesitamos reportar: el valor de F (Fig. 11, celda naranja), los grados de libertad dentro y entre grupos (celda azul, y celda blanca sobre la azul, respectivamente), y el valor p (celda verde). Como ya sabemos, si el valor de p es menor a 0.05, entonces tenemos una diferencia estadísticamente significativa.

	• - ⊘ - ب					×	rot2 - Excel					函	- 0	×
Datei	Start Einfüg	en Seitenlayout F	ormeln Daten Ül	berprüfen /	Ansicht 🖓 W							Anmelo	den 🗸 Freig	eben
Einfüger Zwischena	Calibri	 11 → A[*] A[*] J → □ → □ △ → A → Schriftart □ 	= = • • •	Textumbru Verbinden usrichtung	ch und zentrieren	Standard • 🗣 • %	v 000 €,0 ,00 000 →,0 Formatie	≢ Als Tab gte Als Tab erung * formatie Forma	elle Zellenform ren * tvorlagen	atvorlagen Zel	ūgen × Σ× then × ₩× mat × & ×	Sortieren und S Filtern * A Bearbeiten	Q uchen und uswählen ≁	^
L30	• E	$\times \checkmark f_x =$	WURZEL(0,5*\$K\$24*	*(1/J30+1/K	80))									~
	В	с	D	Е	F	G	н	1	J	к	L	м	N	
15	29,78110219	97,87783554	58,64591905				Gruppen	Anzahl	Summe	Mittelwert	Varianz			
16	34,32295953	125,2355271	57,70478121				Simple	20	591,030667	29,55153335	20,2598395			
17	30,47751346	67,7698579	54,18264885				Complex	20	1714,0823	85,70411497	538,40768			
18	25,66592516	113,9321056	51,03665076				Coda	20	1021,75273	51,08763634	90,4757502			
19	22,5631599	39,19828977	67,40760001											
20	28,80004674	45,31724157	44,4408445											
21	33,8523906	82,65245902	52,39509168				ANOVA							
22							Streuungsursache	dratsummen	iheitsgrade (re Quadratsumme	Prüfgröße (F)	P-Wert	itischer F-We	rt
23							Unterschiede zwi	32101,445	2	16050,72249	74,1780278	1,3676E-16	3,15884272	
24							Innerhalb der Gri	12333,7221	57	216,3810899				
25														
26							Gesamt	44435,1671	59					
27														
28														
29								Dif. Prom	n1	n2	SE	q		
30							simple-Complex	56,1525816	20	20	3,28923312	17,0716333		
31							simple-coda	21,536103	20	20	3,28923312	6,54745414		
32							complex-coda	34,6164786	20	20	3,28923312	10,5241792		
33														
34														
35														
36														-
	vot2	÷						÷ •						Þ
Bereit												<u> </u>	+ +	100 %
Ŧ	H 赵		😰 🚺 🚫								я ^я ^	■ <i>(ii</i> : \$×	16:26 17 07 2018	7

Figura 11: Resultados de la prueba ANOVA.

4.2. Prueba post-hoc: Tukey-Kramer

¿Qué hacemos si la prueba ANOVA es estadísticamente significativa? ¡Recordemos que la prueba ANOVA sólo nos dice que hay una diferencia entre los tres grupos, pero no nos dice entre cuáles! Entonces ahora tenemos que encontrar qué pares son diferentes. Para eso usamos un prueba Tukey-Kramer de diferencias significativas (Ver tabla inferior, Fig. 11). Vamos a dividir el cálculo en dos pasos:

 Calculamos el error estándar de un par, usando esta fórmula. Se puede ver escrita en la barra de fórmulas en la Fig. 11 :

$$SE = \sqrt{\frac{1}{2}MS_w(\frac{1}{n_i} + \frac{1}{n_j})}$$

Donde:

- MS_w es el cuadrado promedio (*Mittlere Quadratsumme*) entre grupos.
- n_i es el número de observaciones (N) de uno de los grupos.
- n_j es el número de observaciones (N) del otro grupo.

• Calculamos el *valor q*, con esta fórmula:

$$q = |M_1 - M_2|/SE$$

...y luego comparamos el valor en una tabla de valores críticos de q (Ver adjunto), buscando en la primera columna los grados de libertad (Fig. celda azul) y en la primera fila el número de grupos. Si nuestro valor q es mayor al valor crítico, entonces el par que analizamos es diferente. Procedemos a hacer lo mismo con todos los pares y así sabremos cuál es el correcto. En este caso, todos los pares son estadísticamente significativos.

Finalmente, reportamos. En el caso de ANOVA, debemos dar los grados de libertad entre grupos (Unterschiede zwischen den Gruppen) y dentro del grupo (Innerhalb der Gruppen).

> Se realizó una prueba ANOVA unifactorial para verificar si existen diferencias entre el VOT de ataques simples, complejos, y en coda. Los resultados muestran que la diferencia es estadísticamente significativa (F= 74.17 (2, 57), p<0.05). Se realizó un test post-hoc Tukey-Kramer, el cual demostró que las tres condiciones presentan valores de VOT diferentes entre sí.

4.3. Graficando resultados

Este caso es exactamente igual que el anterior. Podemos usar el gráfico de cajas, o bien un gráfico de barras. Nuevamente, es importante tener las barras de error. NOTA: también se puede usar el error estándar o la desviación estándar para los gráficos de datos analizados con t-test o ANOVA (no necesariamente el margen de error del 5%). Eso sí, es importante mencionar qué presentan las barras de error.

5. Variables categóricas

Hasta ahora, hemos asumido que la variable dependiente (i.e. la que cambia de acuerdo a una condición) siempre es numérica y continua. Pero qué pasa cuando tenemos simplemente cuentas? Por ejemplo, Si le hemos pedido a un grupo de personas que escuchen dos sonidos y digan si esos sonidos son iguales o diferentes, entonces solo tendremos un número de respuestas "igual" y otro número de respuestas "diferente". En discr.csv tenemos las respuestas igual y diferente de un grupo de hablantes de español que escucharon las vocales [y] y [ø] separadas por un intervalo de 1200 ms (largo), y luego una vez más pero separadas por un intervalo de 300 ms (corto). Queremos saber si este intervalo entre los estímulos (*Interstimulus Interval* o ISI) juega un papel en la discriminación de estos sonidos en hablantes no nativos.

5.1. Cálculo de Chi cuadrado

Para este tipo de datos necesitamos calcular el **Chi cuadrado**. Hacemos con estas respuestas una **tabla de contingencia** como la siguiente (Tabla 1).

Tabla 1: Tabla de contingencia para calcular Chi cuadrado.

	Largo	Corto	Total
Igual	35	15	50
Diferente	5	25	30
Total	40	40	80

Para hacer nuestra tabla, Usaremos la función ZÄHLENWENN. Hacemos clic en una celda vacía (p.ej. en H17) y escribimos en ella lo siguiente:

=ZÄHLENWENN(C2:C41, ''igual'')

Y apretamos Enter, lo cual nos dará el número de respuestas "igual" dadas por los sujetos cuando el ISI es largo (Fig. 6). Ahora, en la celda inmediatamente debajo de donde hicimos esto (H18), escribimos:

=ZÄHLENWENN(C2:C41, ''diferente'')

Y apretamos Enter. Con esto tenemos el número de respuestas "diferente" cuando el ISI es largo. Luego, hacemos clic en la celda inmediatamente debajo (H19) y luego hacemos clic en el botón de sumatoria (Fig. 7). Eso sumará automáticamente las celdas superiores. Con esto tenemos la primera columna de la tabla.

Para generar la columna siguiente, solamente necesitamos repetir el proceso, pero considerando el rango de celdas C42:C81, que es donde están las respuestas con ISI corto.

Ahora, para calcular el Chi cuadrado necesitamos crear una segunda tabla con los **valores esperados** para cada respuesta. Esto podemos hacerlo de la siguiente manera: Seleccionamos cuatro celdas vacías (p.ej. M17, N17, M18 y N18) y vamos a la barra de fórmulas, donde escribimos:

Y luego apretamos Ctrl + Shift + Enter. Esto nos dará una tabla con los valores:

 $\begin{array}{rrr}
 15 & 15 \\
 25 & 25 \\
 \end{array}$

Lo que hicimos es un cálculo de probabilidad si asumimos H_0 , y como podemos ver, este cálculo asume que da igual si el ISI es corto o largo. Podemos también hacer que la

	Largo	Corto	Total
Igual	15	15	30
Diferente	25	25	50
Total	40	40	80

tabla se vea de la siguiente manera (y así sabemos qué números son éstos cuando volvemos a trabajar después de un recreo):

Ahora, por fin, calculamos el Chi cuadrado. Vamos a una celda vacía (p.ej. J23), luego al botón de funciones, y buscamos CHIQU.TEST. En el campo de valores obtenidos (*Beob_ Messwerte*) seleccionamos las celdas correspondientes, y luego hacemos lo mismo en el campo de valores esperados (*Erwart_ Werte*). En la celda aparecerá nuestro valor p, que debería ser: 3.85962E-06 (= $3.85963^{-6} = 0.00000385963$). Claramente, el valor es menor a 0.05, que es el valor p que asumimos como significativo. Por lo tanto, podemos rechazar H₀.

Para reportar esto en el informe, necesitaremos también calcular el valor de Chi cuadrado en sí mismo. Lo hacemos haciendo clic en una celda vacía y escribimos lo siguiente:

=CHIQU.INV.RE(J23,1)

Lo cual nos dará un resultado de 21.333. Pero qué fue lo que acabamos de hacer? Simple: J23 es la celda donde tenemos el valor p. Luego le hemos dado a Excel los grados de libertad, que en este caso es 1 (Cuando tenemos una tabla de contingencia, se calcula: número de columnas -1 * Número de filas -1).

Entonces, el reporte de la prueba se hace así:

Se calculó una prueba de Chi cuadrado comparando las respuestas a estímulos de ISI largo e ISI corto. La prueba arrojó resultados significativos (χ^2 (1) = 21.33, p < 0.05): los pares de estímulos con ISI de 300 ms tienen mayor probabilidad de ser percibidos como diferentes que los pares con ISI de 1200 ms.

5.2. Graficando variables categóricas

La manera más sencilla de graficar este tipo de información es con un gráfico de columnas apiladas Gestapelte Säulen. Simplemente seleccionamos nuestra tabla de contingencia con los valores observados, y luego vamos a Einfügen > Empholene Diagramme > Gestapelte Säule (100 %). Esto nos dará un gráfico prácticamente listo con los valores en forma de porcentajes. Ahora solo tenemos que poner los nombres a los ejes y un título.

6. Algunos datos útiles

- 1. Las funciones en Excel pueden ejecutarse a través de cuadros de diálogo (como lo hicimos con la R de Pearson) o escribiendo directamente la función en la celda (como lo hicimos con Chi cuadrado).
- 2. La versión 2016 de Office tiene todas las funciones estadísticas en el botón Datenanalyse. Esto incluye t-test, Anova, R de Pearson, etc. y además nos da el valor de la prueba estadística, los valores p, los grados de libertad, etc. Es muy útil, pero lamentablemente las versiones anteriores de Excel no lo tienen.

	5	• ~ •	÷		disc	r - Excel			Di	iagrammtool	ls								T	-	٥	×
Dat	ei	Start	Einfügen	Seitenlayout	Formeln	Daten	Überprü	ifen Ansich	nt Entv	vurf For	mat 🤤	Was möd	nten Sie tu	un?					Anm	elden	,Q₁ Freige	ben
Pivot	۲ able E F	mpfohle PivotTabl	ene Tabelle les	Bilder Online	grafiken	🛍 Stor	e ne Add-Ins Add-Ins	s • 👍 Emp Diag	fohlene ramme	2D-Säule			in Ltv	Säule Gewini Verlus Sparklines	n/ Datenschnitt	t Zeitachse	Link Tex	tfeld Kop Fuf Te	f- und Bzeile	πF ΩS Sy	mbole	^
Diag	ramn	n 4 ×	: ×	$\checkmark f_x$						3D-Säule												v
1 2	1 2 3 4 5 6 7 8 9 10	s1 s2 s3 s4 s5 s1 s2 s3 s4	A IS la la la la la la la la la la	B rgo rgo rgo rgo rgo rgo rgo rgo	C RESP igual diferente igual igual igual igual igual igual		D	E 1007 905 argo 805 corto 705	% % % %	2D-Balken			91		J	к •	15	L 25 25	M		N	
-	11 12 13 14 15 16 17	s5 s1 s2 s3 s4 s5 s1	la la la la la la	rgo rgo rgo rgo rgo rgo	igual igual igual igual igual igual igual			605 505 405 305 205 105 05	% % % % %	M Weitere	e Säulenc	liagramme.					difere	ente	largo	15	orto	15
	18 19 20 21 22 23	s2 s3 s4 s5 s1 s2	la la la la la	rgo rgo rgo rgo rgo	igual igual igual igual igual			0		di	iterente	largo	corto	igual	#NAME?	0	igual			25 40		25 40
		She	et1 (÷																		•
Bereit	Berec	hnen												Mittelwert: 20	Anzahl:	8 Summe:	80 🏢				+	100 %
		Ξi	6	w I	P	×∃	1										RR	~ 4	∎ <i>(</i> ~ \$×	13.07	.2018	1

Figura 12: Gráfico de barras apilado en porcentajes.

- 3. Si usas Excel en Windows, se puede descargar el add-on Analysis Toolpak (*Analyse-Funktionen*). Lamentablemente, esto no está disponible para Excel Mac 2011.
- 4. Las pruebas estadísticas t-test y Anova asumen un par de supuestos que no hemos analizado aquí:
 - ¿Tienen los datos una distribución normal?
 - ¿Tienen los datos igual varianza?

...pero esto lo veremos caso por caso.

- 5. En estos tiempos donde todo está en Internet, podemos googlear algo como wie kann ich ein Boxplotdiagramm in Excel erstellen y encontraremos cientos de páginas, tutoriales en Youtube, etc.
- 6. *R* es un programa de libre acceso (open source) y aunque parece complicado, hace las cosas muy simples una vez que entendemos cómo funciona. Si tienes tiempo e interés en el análisis de datos, esta alternativa es siempre la mejor. Otras alternativas (pagada\$\$\$) son SPSS y SAS.
- 7. El gráfico de cajas era lo ideal para los datos en vot.csv. Si de verdad deseas hacerlo en una versión de Excel anterior a 2016, aquí hay un buen tutorial:

https://www.youtube.com/watch?v=jaXwtPKUGL8