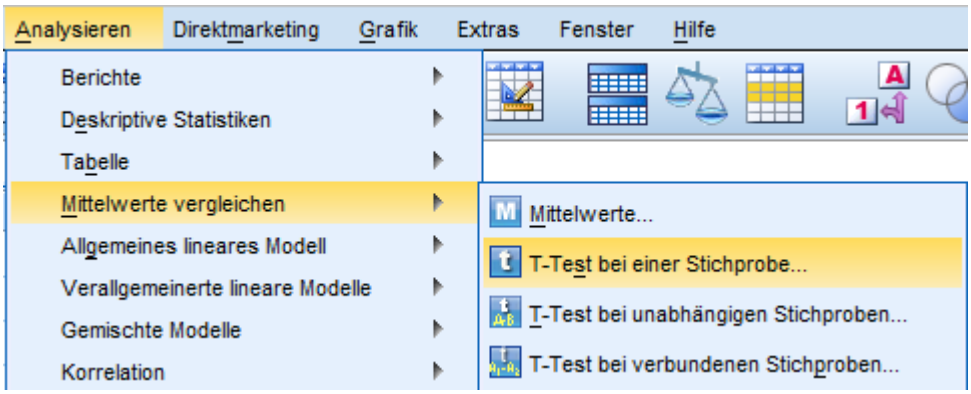
1.1. Vergleich von Mittelwerten

1.1.1. Eine glückliche Familie: Die t-Tests

Die einfachsten inferenzstatistischen Verfahren sind die *t*-Tests. Bei ihnen dreht sich alles um Mittelwerte und die Frage: „Sind diese beiden Werte gleich oder ungleich?“

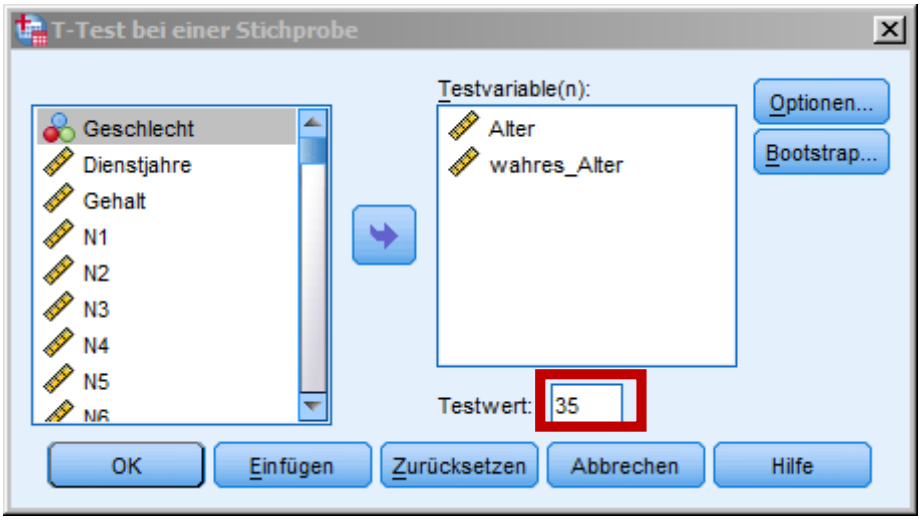
1.1.1.1. Der t-Test für eine Stichprobe

Beim *t*-Test für eine Stichprobe (auch Ein-Stichproben *t*-Test, one sample *t*-test) wird getestet, ob der Mittelwert einer Variablen einem bekannten festen Wert entspricht. Ted hat irgendwo gelesen, dass Krankenhausmitarbeiter im Mittel 35 Jahre alt sind. Er möchte nun überprüfen, ob das auch im Sacred Heart so ist. Im Kapitel zum Umkodieren mit Bedingung hatte er festgestellt, dass einige Mitarbeiter bei ihrer Altersangabe geschummelt haben. Wir möchten daher sowohl die Variable Alter, als auch die Variable wahres_Alter überprüfen. Über „Analysieren“ → „Mittelwerte



vergleichen“ gelangen wir zu den *t*-Tests und wählen „T-Test bei einer Stichprobe“ (SPSS schreibt immer ein großes T, es sollte aber eigentlich ein kleines sein).

Im folgenden Dialog wählen wir alle Variablen aus, die wir gegen seinen festen Wert (die 35) testen wollen.



Die eingefügte Syntax sieht dann so aus:

```
DATASET ACTIVATE DataSet1.
T-TEST
  /TESTVAL=35
  /MISSING=ANALYSIS
  /VARIABLES=Alter wahres_Alter
  /CRITERIA=CI(.95).
```

Die Zeile mit **DATASET ACTIVATE** kann er wie immer löschen. Die Ausgabe sieht dann so aus:

Statistik bei einer Stichprobe

	H	Mittelwert	Standardabweichung	Standardfehler Mittelwert
Alter	328	33.78	8.771	.484
wahres_Alter	328	35.2134	8.74985	.48313

Test bei einer Stichprobe

	Testwert = 35					
					95% Konfidenzintervall der Differenz	
	t	df	Sig. (2-seitig)	Mittelwertdifferenz	Unterer	Oberer
Alter	-2.518	327	.012	-1.220	-2.17	-.27
wahres_Alter	.442	327	.659	.21341	-.7370	1.1638

Die erste Tabelle enthält noch einmal die deskriptiven Kennwerte der getesteten Variablen. In der zweiten Tabelle schauen wir auf die Signifikanz: Sie liegt beim Alter bei .012, beim wahren Alter bei .659. Wie üblich vergleichen wir mit dem α-Niveau von 5%: Der Test wird also für das Alter signifikant, für das wahre Alter allerdings nicht.

Da wir niemals nur auf die Signifikanz sondern immer auch auf die Effektstärke schauen sollten, berechnen wir nun noch Cohens d. Das macht SPSS leider nicht für uns. Wir rechnen daher von Hand (oder auch per Excel oder Taschenrechnerfunktion):

$$d = \frac{|\bar{x} - \mu|}{\sigma}$$

Für das Alter erhalten wir:

$$d_{Alter} = \frac{|33.78 - 35|}{8.771} = 0.139$$

Für das wahre Alter erhalten wir:

$$d_{wahr} = \frac{|35.2134 - 35|}{8.74985} = 0.024$$

Ted schreibt daher folgendes in seinen Bericht:

> „Um zu prüfen, ob die angegebenen Alterswerte mit einem theoretischen Mittelwertwert von 35 vereinbar sind, wurden Ein-Stichproben *t*-Tests für die Variablen Alter und Wahres Alter durchgeführt. Der Mittelwert des Alters war signifikant verschieden von 35, $t(327) = -2.518$, $p = .012$. Die Effektgröße *d* ist allerdings nach Cohen nur als sehr klein einzuordnen ($d = 0.139$). Wir fanden keine signifikante Abweichung des wahren Alters von 35, $t(327) = 0.442$, $p = .659$, $d = 0.024$."

1.1.1.2. Der t-Test für abhängige Stichproben

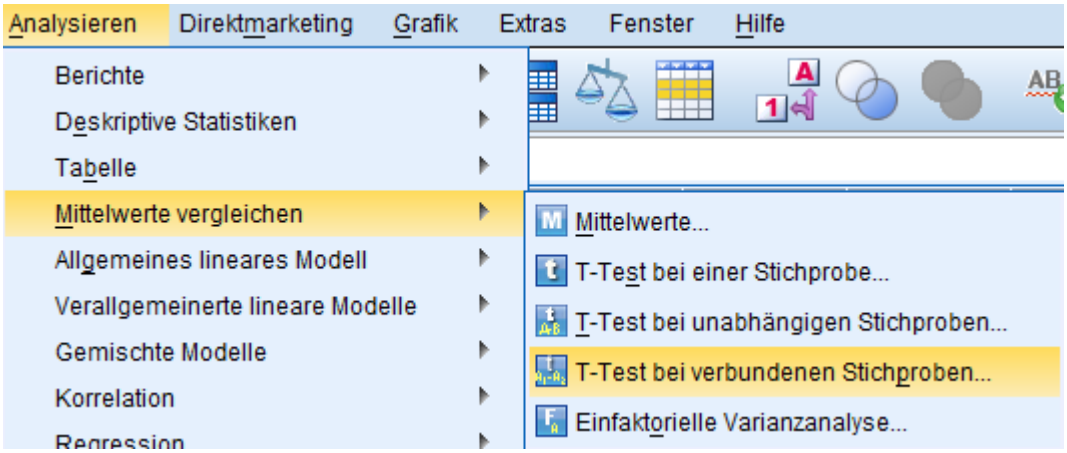
Unter den Männern im Sacred Heart ist ein erbitterter Männlichkeitswettstreit entbrannt. Dr. Cox hat jedem Mann 7 ManCards gegeben (ManCards_T1). Für besonders männliches Verhalten erhält man eine Karte dazu, für besonders unmännliches Verhalten (Komplimente über Schuhe und Handschriften, das Einnehmen der Anti-Baby-Pille) verliert man eine Karte. Dr. Cox kontrolliert täglich, wie viele Karten jeder Mitarbeiter noch übrig hat. In extremen Fällen kann es sogar dazu kommen, dass ein Mitarbeiter in den negativen Bereich rutscht (wer keine ManCard mehr hat, aufgrund unmännlichen Verhaltens aber eine verlieren müsste, erhält stattdessen eine GirlCard [als negativer Wert in der ManCard-Variable notiert). Ted ist daran interessiert, ob sich die Anzahl der ManCards pro Person von Tag 2 zu Tag 3 und von Tag 3 zu Tag 4 verändert.

Es geht also jeweils um den Vergleich zweier Mittelwerte, wobei aber in beiden Mittelwerten Werte derselben Personen stecken. Wir haben demnach eine Messwiederholung. Hier wird daher der t-Test für abhängige Stichproben verwendet[1].

Zunächst muss Ted alle Frauen aus dem Datensatz herausfiltern, sonst würden ihre 0-Werte (da sie ja keine ManCards erhalten) sinnloser Weise in die Berechnung mit einfließen:

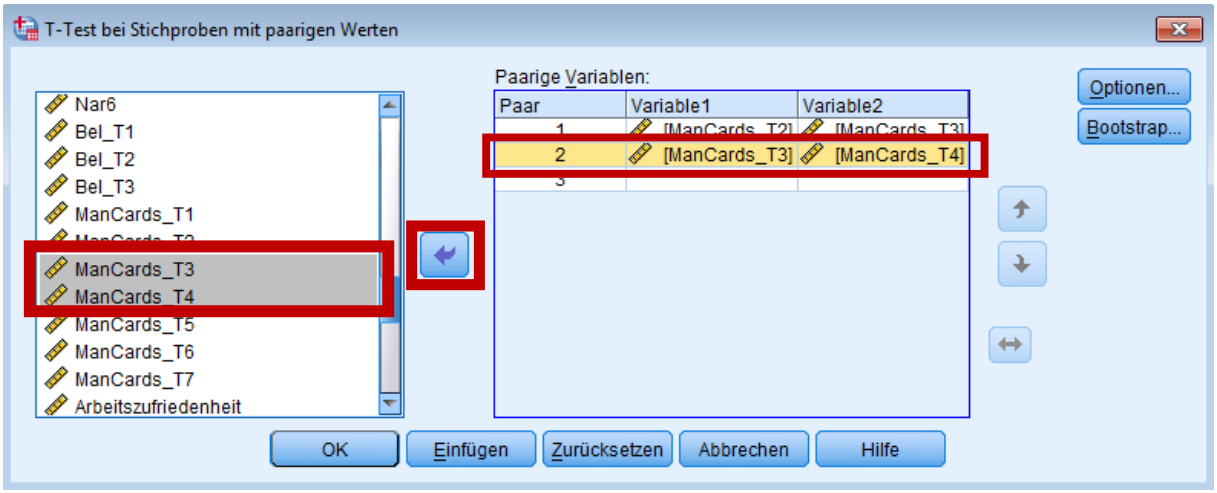
```
COMPUTE Filter1 = (Geschlecht = 1).
EXECUTE.
FILTER BY Filter1.
```

Den t-Test für abhängige Stichproben findet er unter „Analysieren" → „Mittelwerte vergleichen" → „T-Test bei verbundenen Stichproben…":



[1] Für Kenner: Natürlich wäre eine noch schönere Variante die rmANOVA, die wird aber in einem anderen Kapitel besprochen.

Im folgenden Dialogfeld kann er nun theoretisch beliebig viele abhängige t-Tests anfordern: Für jeden Test muss er nur die beiden Variablen, die er gegeneinander testen möchte, als „Paar“ angeben. In diesem Falle gibt er also ManCards_T2 und ManCards_T3 für den ersten Test (von Tag 2



gegen Tag 3) an, ManCards_T3 und ManCards_T4 für den zweiten Test:

Hier braucht er keine weiteren Angaben zu machen. Er fügt einfach nur die Syntax ein:

```
T-TEST PAIRS=ManCards_T2 ManCards_T3 WITH ManCards_T3 ManCards_T4 (PAIRED)
  /CRITERIA=CI(.9500)
  /MISSING=ANALYSIS.
```

Die Ausgabe sieht dann aus wie folgt. Wie gewohnt bekommen wir zunächst eine Tabelle der Deskriptiven Statistik:

Statistik für Stichproben mit paarigen Werten

		Mittelwert	H	Standardabweichung	Standardfehler Mittelwert
Paar 1	ManCards_T2	6,06	93	2,146	,223
	ManCards_T3	5,34	93	3,077	,319
Paar 2	ManCards_T3	5,34	93	3,077	,319
	ManCards_T4	4,47	93	3,933	,408

Hier kann er bereits ablesen, dass deskriptiv die Mittelwert immer weiter abnehmen. Die eigentlichen t-Tests findet er in folgender Tabelle:

Test für Stichproben mit paarigen Werten

		Paarige Differenzen					t	df	Sig. (2-seitig)
		Mittelwert	Standardabweichung	Standardfehler Mittelwert	95% Konfidenzintervall der Differenz Unterer	Oberer			
Paar 1	ManCards_T2 - ManCards_T3	,720	2,108	,219	,286	1,155	3,296	92	,001
Paar 2	ManCards_T3 - ManCards_T4	,871	2,023	,210	,454	1,288	4,152	92	,000

Der eigentliche t-Test (mit t-Wert, Freiheitsgraden und p) steht ganz rechts außen. Hier stellen wir fest: Beide t-Tests werden signifikant, die Anzahl der ManCards nimmt also kontinuierlich ab.

Selbstverständlich wollen wir aber auch hier ein Maß der Effektstärke berechnen. Wie wir wissen, macht der abhängige t-Test nicht anderes, als die Differenzdatenreihe der beiden Messzeitpunkten zu bilden und diese dann als Einstichproben-t-Test gegen 0 zu testen. Die deskriptiven Eigenschaften dieser Differenzvariablen finden wir im linken Teil der Tabelle. Wir können also wieder die bereits bekannte Formel für Cohen´s d verwenden und erhalten:

Für T2 gegen T3:

$$d_{T2-T3} = \frac{|.724 - 0|}{2.108} = 0.343$$

Und für T3 gegen T4:

$$d_{T3-T4} = \frac{|.871 - 0|}{2.023} = 0.431$$

Ted schreibt daher folgendes in seinen Bericht:

> Um zu überprüfen, ob sich die durchschnittliche Anzahl von ManCards pro Person von Tag 2 zu Tag 3 und von Tag 3 zu Tag 4 verändert, wurden *t*-Test für abhängige Stichproben der jeweiligen Tage gegeneinander berechnet. Für den Test von Tag 2 gegen Tag3 ergab sich ein signifikanter Abfall von 0.72 Karten, $t(92)=3.30$, $p=.001$, $d=0.34$. Dieser Effekt ist als klein einzustufen. Für den Test von Tag 3 gegen Tag 4, ergab sich ein leicht stärkerer Abfall von 0.87 Karten, $t(92)=4.152$, $p<.001$, $d=0.43$. Dieser Effekt ist als klein bis mittelstark einzustufen.
> Zusammenfassend lässt sich sagen, dass die Anzahl von ManCards kontinuierlich abnahm, im zweiten Erhebungszeitraum etwas stärker als im ersten.

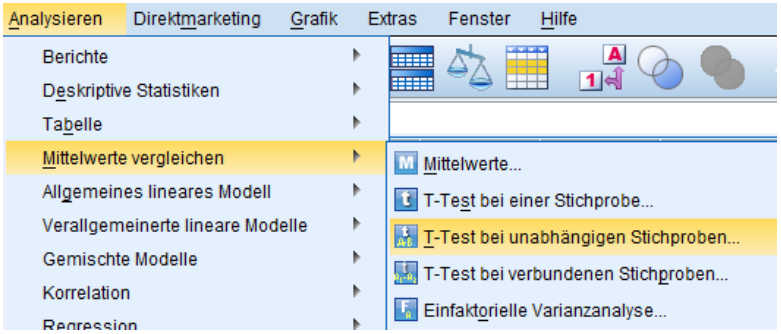
Anschließend schaltet er den Filter wieder aus:

```
FILTER OFF.
```

1.1.1.3. Der t-Test für unabhängige Stichproben

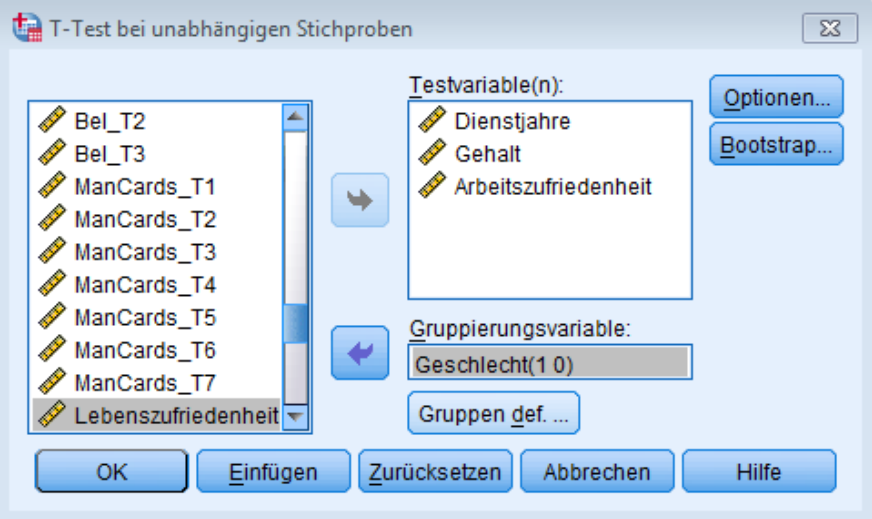
Die letzte Variante des t-Tests ist der t-Test für unabhängige Stichproben. Hier vergleichen wir, ob zwei Mittelwerte ungleich sind, wobei jede Person nur in einen der beiden Mittelwerte mit einem Wert eingeht.

Ted soll prüfen ob es einen Geschlechterunterschied bezüglich der Dienstjahre, des Gehalt und der Arbeitszufriedenheit gibt. Anders formuliert: Haben Männer und Frauen unterschiedliche Mittelwerte bei Dienstjahren, Gehalt und Arbeitszufriedenheit?

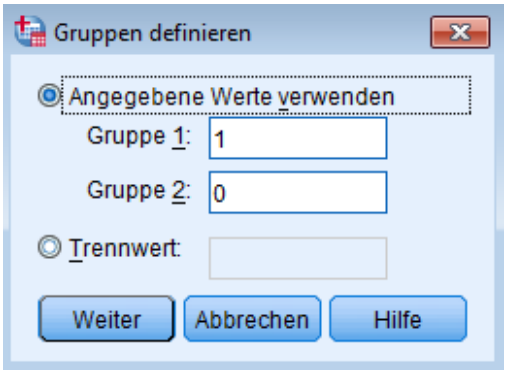


Er findet den unabhängigen t-Test unter „Analysieren“ → „Mittelwerte vergleichen“ → „T-Test bei unabhängigen Stichproben…“.

Im folgenden Dialogfeld kann er beliebig viele t-Test gleichzeitig anfordern, solange sie nach den selben Gruppen aufgeteilt werden sollen. Dazu zieht er alle Variablen, aus denen Mittelwerte gebildet werden sollen (in der Regel sind dies die Abhängigen Variablen), in das Feld „Testvariable(n)“ und die Gruppierungsvariable (überraschender Weise) in das Feld „Gruppierungsvariable“.



Unter „Gruppen def.“ (definieren) muss er noch angeben, welche Werte die die zu vergleichenden Gruppen haben: Die Variable „Geschlecht“ enthält die Werte 0 (weiblich) und 1 (männlich), daher trägt er diese Werte bei Gruppe 1 und Gruppe 2 ein:



Theoretisch kann man dies auch mit String-Variablen machen: Dann würde man die Zeichenfolge (das Wort) das die jeweilige Gruppe bezeichnet, in das jeweilige Feld eintragen. Molly warnt aber davor, dass SPSS bei String-Variablen im unabhängigen t-Test ab und zu herumzickt. In diesem Falle sollte man die Variable am besten in eine numerische Variable umkodieren.

Mehr braucht Ted nicht auszuwählen, daher fügt er den Test in die Syntax ein:

```
T-TEST GROUPS=Geschlecht(1 0)
  /MISSING=ANALYSIS
  /VARIABLES= Dienstjahre Gehalt Arbeitszufriedenheit
  /CRITERIA=CI(.95).
```

Zunächst erhält er wie immer eine Tabelle mit deskriptiven Kennwerten:

Gruppenstatistik

	Geschlecht	H	Mittelwert	Standardabweichung	Standardfehler Mittelwert
Dienstjahre	1	93	14,56	10,610	1,100
	0	235	13,23	8,252	,538
Gehalt	1	93	3170,16	3188,200	330,601
	0	235	2266,63	2351,556	153,399
Arbeitszufriedenheit	1	93	6,00	2,493	,259
	0	235	5,05	2,317	,151

Hier schaut er wie immer nur kurz auf die Mittelwerte um einen Hinweis darauf zu bekommen, in welche Richtung die Effekte gehen könnten.

Von der nächsten Tabelle schaut er sich zunächst nur die linke Hälfte an:

		Levene-Test der Varianzgleichheit	
		F	Sig.
Dienstjahre	Varianzgleichheit angenommen	4,214	,041
	Varianzgleichheit nicht angenommen		
Gehalt	Varianzgleichheit angenommen	8,794	,003
	Varianzgleichheit nicht angenommen		
Arbeitszufriedenheit	Varianzgleichheit angenommen	,464	,496
	Varianzgleichheit nicht angenommen		

Diese linke Hälfte enthält bereits einen Signifikanztest, der aber mit dem eigentlichen Mittelwertsunterschied nicht zu tun hat: Der **Levene-Test** überprüft, ob die Daten in beiden Gruppen die gleiche (bzw. eine hinreichend ähnliche) Varianz haben[2]. Da es sich hier um einen Homogenitätstest handelt, testen wir nicht auf den üblichen 5% Signifikanzniveau, **sondern auf 25%** (dadurch wird der Test früher signifikant und warnt uns frühzeitig, falls unsere Varianzen inhomogen sind). Ted stellt fest, dass bei den Variablen Dienstjahre und Gehalt die **Varianzhomogenität** verletzt ist (die p-Werte des Levene-Tests liegen unter 25%). Bei diesen beiden Variablen dürfen wir also dem klassischen t-Test eigentlich nicht vertrauen. Ted behält das im Hinterkopf und sieht sich die rechte Hälfte der Tabelle an:

T-Test für die Mittelwertgleichheit						
					95% Konfidenzintervall der Differenz	
t	df	Sig. (2-seitig)	Mittelwertdifferenz	Standardfehlerdifferenz	Unterer	Oberer
1,212	326	,226	1,334	1,100	-,831	3,498
1,089	138,206	,278	1,334	1,225	-1,088	3,755
2,820	326	,005	903,532	320,346	273,325	1533,738
2,479	133,447	,014	903,532	364,456	182,674	1624,389
3,285	326	,001	,953	,290	,382	1,524
3,183	158,362	,002	,953	,300	,362	1,545

[2] Für Kenner der ANOVA: Das ist die gleiche Voraussetzung, die wir auch bei den ANOVAs ohne Messwiederholung testen. Die Auswahl des zu interpretierenden p-Wertes entspricht dem Vorgehen bei der einfaktoriellen ANOVA: Die Entscheidung zwischen ANOVA und Brown-Forsythe entspricht der Entscheidung zwischen t-Test und Welch-Test.

Hier werden uns pro Variable zwei Signifikanzwerte ausgegeben:

Der obere (z.B. .001 für die Variable Arbeitszufriedenheit) ist der p-Wert des **klassischen t-Tests**. Ihn interpretieren wir immer dann, wenn wir homogene Varianzen haben. Im konkreten Beispiel dürfen wir das also nur für die Variable Arbeitszufriedenheit (bei den anderen beiden war der Levene-Test ja signifikant geworden).

Der untere Signifikanzwert (.278 für Dienstjahre und .014 für das Gehalt) ist der p-Wert des so genannten **Welch-Tests**. Dieser ist quasi ein t-Test, der aber nicht von Varianzhomogenität ausgeht. Ihn interpretieren wir daher dann, wenn die Varianzhomogenität verletzt ist (der Levene-Test also signifikant geworden ist).

Nun brauchen wir nur wieder Cohen's d als Maß der Effektstärke. Leider ist es beim unabhängigen t-Test etwas aufwendiger zu berechnen als bei den vorangegangenen t-Tests:

$$d = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{(n_1 * {SD_1}^2 + n_2 * {SD_2}^2)}{n_1 + n_2 - 2}}} = \frac{|\bar{x}_1 - \bar{x}_2|}{SD_{pooled}}$$

Dabei sind $\bar{x}_1$ und $\bar{x}_2$ die Mittelwerte der beiden Gruppen, n_1 und n_2 die Gruppengrößen und SD_1 und SD_2 die Standardabweichungen. Ted ist so erschlagen von der Formel, dass er sie nur für eine der Variablen, die Arbeitszufriedenheit, ausrechnet:

$$d_{Arb.Zufr.} = \frac{|6.00 - 5.05|}{\sqrt{\frac{(93 * 2.493^2 + 235 * 2.317^2)}{93 + 235 - 2}}} = \frac{0.95}{2.375} = 0.399$$

Glücklicherweise hat Molly noch einen Trick im Ärmel: Wenn man den t-Wert des t-Tests bereits kennt, gibt es eine vereinfachte Berechnungsformel:

$$d = t * \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Mit dieser vereinfachten Formel ergibt sich für die beiden übrigen Variablen[3]:

$$d_{Dienstjahre} = 1.212 * \sqrt{\frac{1}{93} + \frac{1}{235}} = 0.148$$

$$d_{Gehalt} = 2.820 * \sqrt{\frac{1}{93} + \frac{1}{235}} = 0.345$$

[3] Auch wenn wir den Welch-Test interpretieren, verwenden wir den t-Wert des klassischen t-Tests zur Berechnung von Cohen's d.

Nun schreibt Ted wie immer seine Interpretation:

> Die Geschlechtsunterschiede in Bezug auf die geleisteten Dienstjahre, das Gehalt und die Arbeitszufriedenheit wurden mittels t-Tests für unabhängige Stichproben untersucht. Wenn der Levene-Test auf Varianzheterogenität hinwies, wurde stattdessen der Welch-Test interpretiert.
> Männer und Frauen unterscheiden sich nicht signifikant bezüglich ihrer Dienstjahre, $t_{Welch}(138.21)=1.09$, $p=.278$, $d=0.15$. Für das Gehalt ergab sich ein signifikanter Unterschied: Männer verdienen im Schnitt 900€ mehr, $t_{Welch}(133.45)=2.48$, $p=.014$, $d=0.35$. Dieser Unterschied ist als klein bis mittelstark einzuschätzen. Auch für die Arbeitszufriedenheit ergab sich ein signifikanter Unterschied: Männer sind im Schnitt einen Punkt zufriedener mit ihrer Arbeit, $t(326)=3.29$, $p=.001$, $d=0.40$. Dieser Effekt ist als klein bis mittel einzuschätzen.