Kommentar: Falls Sie das lesen, erstmals: Herzliche Gratulation! Viele Leute sind leider schon lange abgehängt worden. Es ist auch völlig normal, dass man bei Hypothesentests abgehängt wird. Falls das mit Ihnen passiert ist, machen Sie sich keine Sorgen - wir können das Thema wiederholen. Und ich habe Ihnen die Theorie zu Hypothesentests im Anschluss an die Übungen noch aufgeschrieben. Vielleicht hilft das. Wichtig in diesem Zusammenhang ist, dass man dran bleibt, die Geduld nicht verliert und nicht aufgibt. Der Endspurt ist immer anstrengend, aber es geht wirklich nicht mehr lange. *The hitchhiker's guide to the galaxy* sagt das ganz schön mit

Aufgabe 1 Dieses Beispiel haben wir im Unterricht bereits angetroffen; wir haben aber zweiseitige Tests ausgeführt (siehe Jupyter Notebook). Jetzt geht es darum, dass Sie einseitig testen. Hier nochmals die Aufgabenstellung.

Der Müesli-Produzent hat die Vermutung, seine Maschine sei falsch kalibriert und packe zu viel Müesli in die Säcke, was seine Marge drückt. Um seiner Vermutung nachzugehen, wiegt er den Inhalt von 50 zufällig ausgewählten Müeslipackungen. Die Stichprobe finden Sie auf Moodle im File `Müesli example (a new sample)`.

(a) Führen Sie einen geeigneten einseitigen z-Test von Hand mit Hilfe der Normalverteilungstabelle durch. Schreiben Sie das Testergebnis inkl. alle Schritte auf.

(b) Führen Sie einen geeigneten einseitigen t-Test mit Python durch. Schreiben Sie das Testergebnis inkl. alle Schritte auf.

Aufgabe 2 Ein Investitionsschema verspricht variable monatliche Renditen. Ein Investor wird nur dann darin investieren, wenn ihm ein durchschnittliches Monatseinkommen von 180 CHF zugesichert wird. Er hat eine Stichprobe von 300 Monatsrenditen gemessen, die einen Mittelwert von 190 CHF und eine Standardabweichung von 75 CHF aufweist. Sollte er oder sie in dieses System investieren?

(a) Führen Sie einen z-Test von Hand mit Hilfe der Normalverteilungstabelle durch. Schreiben Sie das Testergebnis inkl. alle Schritte auf.

(b) Führen Sie einen t-Test mit Python durch. Schreiben Sie das Testergebnis inkl. alle Schritte auf.

Aufgabe 3 Die Theorie für Teilaufgabe (c) haben wir noch nicht vollständig behandelt; sie finden diese aber weiter unten im Dokument im Abschnitt *Der Einstichproben-t-Test*. Probieren Sie die Teilaufgabe (c) mit Hilfe der Theore zu lösen. Falls es klappt, ist es toll! Und ansonsten denken Sie daran „DON'T PANIK". Wir behandeln die Theorie im Unterricht.

Wir vermuten, dass die mittlere Grösse von Erwachsenen in der Schweiz kleiner gleich 170 cm ist. Nun möchten wir diese Annahme überprüfen und messen dafür die Ihre Grösse (d.h. die Grösse der Statistik-Studierenden vom FS 2020). (siehe File `Stats_Height_Weight.xlsx` auf Moodle für die Messwerte).

(a) Die Stichprobengrösse beträgt nur $n = 23 (< 30)$ und wir somit dürften hier eigentlich keinen z-Test ausführen. Warum ist das so? Führen Sie dennoch einen geeigneten z-Test von Hand mit Hilfe der Normalverteilungstabelle durch, damit Sie das üben können.

(b) Führen Sie einen geeigneten t-Test mit Python durch.

(c) Kontrollieren Sie mit Python, ob die nötigen Voraussetzungen für den t-Test erfüllt werden.

Theorieteil: Der Einstichproben-z-Test

Ein konkretes Beispiel

Ein Müesli-Produzent verwendet eine Abfüllmaschine, welche 500 g Müesli in Säcke abfüllen soll. Er hat die Vermutung, die Maschine sei falsch kalibriert und packe zu viel Müesli in die Säcke, was seine Marge drückt. Um seiner Vermutung nachzugehen, wiegt er den Inhalt von 50 zufällig ausgewählten Müeslipackungen; er erhält folgende Werte:

Messung	1	2	3	4	5	6	7	8	9	$\cdots$	49	50
Masse [g]	514	510	497	508	496	517	503	504	498	$\cdots$	506	510

Gewisse Messwerte liegen unter 500 g, andere darüber. Für die Stichprobe von $n = 50$ Messungen, haben wir einen Mittelwert von $\overline{x} = 507.1$ und eine empirische Standardabweichung von $s = 7.3$ erhalten. Können wir aufgrund der Daten sagen, ob die Abfüllmaschine falsch kalibriert ist? Wie?

Das Beispiel ist eine typische Fragestellung, die sich mit Hilfe eines *statistischen Hypothesentests* beantworten lässt. Eines gleich vorneweg: Wir können auf die Frage oben keine sichere „ja/nein"-Antwort erwarten. Die Statistik (und die zu Grunde liegende Wahrscheinlichkeitsrechnung) helfen uns aber abzuschätzen, ob die gemessenen Müesli-Massen unter der Annahme, dass die Abfüllmaschine *korrekt* arbeitet, *plausibel* sind.

Das Verfahren in 7 Schritten (wichtig für Sie!!!)

Ein statistischer Hypothesentest folgt immer demselben Ablauf in **7 Schritten**. Wir skizzieren ihn hier am Beispiel der Abfüllmaschine:

1) **Modell** für Daten aufstellen: Bei der Abfüllmaschine dürfen wir erwarten, dass die Messungen unabhängig sind und alle der gleichen Verteilung folgen. Formal: $X_1, X_2, \ldots, X_n$ sind i.i.d. mit Erwartungswert μ und Varianz σ^2 (beide unbekannt).

   Über die konkrete Verteilung der einzelnen Messungen müssen wir hier keine Annahme treffen, da wir mit $n = 50$ eine hinreichend grosse Anzahl von Stichprobenwerten haben, um mit Hilfe des zentralen Grenzwertsatzes (Normalverteilung des Mittelwerts) zu argumentieren. Wir nehmen also insbesondere *nicht* an, dass die einzelnen Messungen normalverteilt sind!

2) Die **Nullhypothese** ist eine Annahme an die Parameter des Modells, die wir anzweifeln, die wir aber benutzen, um die „Plausibilität" der gemessenen Daten zu bestimmen. In unserem Beispiel ist die Nullhypothese, dass die Abfüllmaschine *korrekt* arbeitet, also tatsächlich im Mittel 500 g pro Packung abfüllt. Formal notieren wir das so: $H_0 : \mu = \mu_0 = 500$.

   Als Negation der Nullhypothese ergibt sich die **Alternativhypothese**. Im Beispiel ist die Alternativhypothese offensichtlich $H_\mathrm{A} : \mu \neq \mu_0 = 500$.

   Wir interessieren uns hier also gleichermassen für die Fälle, dass die Abfüllmaschine zu viel oder zu wenig abpackt, und führen deshalb einen sogenannten *zweiseitigen* Test durch. In gewissen Fällen reicht es aber, *einseitig* zu testen. Den Müesli-Produzenten beispielsweise interessiert vor allem, ob seine Abfüllmaschine nicht *zu viel* Müesli abpackt, da das auf seine Marge drückt. Der Konsumentenschutz auf der andern Seite will vor allem wissen, ob die Abfüllmaschine nicht *zu wenig* Müesli abpackt.

Test	Nullhypothese	Alternativhypothese
zweiseitig	$H_0 : \mu = \mu_0$	$H_\mathrm{A} : \mu \neq \mu_0$
einseitig (*upper tail*)	$H_0 : \mu \leq \mu_0$	$H_\mathrm{A} : \mu > \mu_0$
einseitig (*lower tail*)	$H_0 : \mu \geq \mu_0$	$H_\mathrm{A} : \mu < \mu_0$

3) Die **Teststatistik** ist eine einzelne Grösse, die wir mit Hilfe der gemessenen Daten berechnen, um ihre Plausibilität abzuschätzen. Beim Problem der Abfüllmaschine scheint intuitiv klar, dass die Teststatistik das Stichprobenmittel $\overline{X}$ der Messwerte verwenden muss, weil dieses als Schätzer für den Erwartungswert μ dient (für den wir ja eine Null- und eine Alternativhypothese aufgestellt haben). Als Teststatistik verwendet man die standardisierte Grösse

$$Z = \frac{\overline{X} - \mu_0}{S/\sqrt{n}} \ .$$

   (Mit S bezeichnen wir hier das zufällige Resultat der empirischen Standardabweichung s. Dieses ist genau so eine Zufallsvariable wie der Mittelwert $\overline{X}$.)

   Unter der Nullhypothese ist Z näherungsweise standardnormalverteilt: Mit $n = 50$ haben wir hinreichend viele Messungen, dass a) laut zentralem Grenzwertsatz näherungsweise eine Normalverteilung resultiert, und b) die empirische Standardabweichung S ein vernünftig guter Schätzer für σ ist, wir bei der Standardisierung also keinen grossen Fehler begehen.

   Für unsere Stichprobe von $n = 50$ Messungen nehmen wir an, dass wir einen Mittelwert von $\overline{x} = 507.1$ und eine empirische Standardabweichung von $s = 7.3$ erhalten. Daraus resultiert ein Wert von $z \approx 6.877$ für die Teststatistik Z.

4) Der ***p*-Wert** bezeichnet die Wahrscheinlichkeit, *unter der Nullhypothese* einen Wert der Teststatistik zu erhalten, der *mindestens so extrem* ist, wie der tatsächlich gemessene. Was „mindestens so extrem“ heisst, hängt von der Alternativhypothese ab. Da wir sowohl $\mu < 500$ als auch $\mu > 500$ als Abweichung von der Nullhypothese werten, ist der p-Wert gegeben durch die Wahrscheinlichkeit

$$P(|Z| \geq 6.877) = P(Z \geq 6.877) + P(Z \leq -6.877) = 2 \cdot (1 - P(Z \leq 6.877))$$

unter der Annahme, dass Z standardnormalverteilt ist.

Mit Python erhalten wir diese Wahrscheinlichkeit wie folgt:

```
> from scipy import stats
> 2*(1 - stats.norm.cdf(6.877, 0, 1))
ans = 6.1127e-12
```

In der folgenden Tabelle bezeichnet Z die *Zufallsvariable* der Teststatistik, z den *Wert*, den sie für einen bestimmten Datensatz annimmt. Diese Formeln gelten für beliebige Tests mit symmetrischer stetiger Teststatistik, also insbesondere für den Einstichproben-z-Test und Tests mit normalverteilter Teststatistik.

Test	Nullh.	Alternativh.	p-Wert
zweiseitig	$H_0 : \mu = \mu_0$	$H_\text{A} : \mu \neq \mu_0$	$p = P(Z \leq -\lvert z\rvert) + P(Z \geq \lvert z\rvert)$ $= 2 \cdot (1 - P(Z \leq \lvert z\rvert))$
einseitig (*upper tail*)	$H_0 : \mu \leq \mu_0$	$H_\text{A} : \mu > \mu_0$	$p = P(Z \geq z) = 1 - P(Z \leq z)$
einseitig (*lower tail*)	$H_0 : \mu \geq \mu_0$	$H_\text{A} : \mu < \mu_0$	$p = P(Z \leq z)$

5) Das **Signifikanzniveau** α ist der kleinste p-Wert, bei dem wir die Daten noch als vereinbar mit der Nullhypothese akzeptieren. Die Wahl des Signifikanzniveaus hängt von der Anwendung ab; für viele Anwendungen wählt man ein Signifikanzniveau von 5%.

6) **Konfidenzintervalle** liefern uns eine Methode, die Unsicherheit eines Punktschätzers anzugeben. Mit Hilfe der Stichprobe können wir nämlich nicht bloss den Punktschätzer berechnen, sondern auch einen Bereich (eben das Konfidenzintervall oder *Vertrauensintervall*), in welchem die unbekannte (wahre) Durchschnittsgrösse μ mit einer gewissen Sicherheit liegt. Diese „Sicherheit“ wird mit *Konfidenzniveau* bezeichnet. Mit Hilfe des zentralen Grenzwertsatzes können wir ein Konfidenzintervall zu einem gegebenen Konfidenzniveau $1 - \alpha$ berechnen.

Das $1 - \alpha = 0.95$ Konfidenzinterval für den **zweiseitigen Test**, kann man am einfachsten mit der folgenden Formel berechnen

$$\left[x_{lower\_limit}\,,\, x_{upper\_limit}\right] = \left[\overline{x} - z_{0.975} \cdot \frac{s}{\sqrt{n}}\,,\; \overline{x} + z_{0.975} \cdot \frac{s}{\sqrt{n}}\right]$$

Wir brauchen nur das $z_{0.975}$ Quantil mit der Tabelle zu bestimmen.

Für den **einseitigen (*upper tail*) Test**, müssen wir verstehen, dass die obere Grenze bereits gegeben ist! Diese ist nämlich ∞. Also müssen wir nur noch die Untergrenze bestimmen.

$$\left[x_{lower\_limit}\,,\, x_{upper\_limit}\right] = \left[\overline{x} - z_{0.95} \cdot \frac{s}{\sqrt{n}}\,,\ \infty\right)$$

Für den **einseitigen (*lower tail*) Test**, müssen wir verstehen, dass die untere Grenze $-\infty$ beträgt. Also müssen wir nur noch die Oberegrenze bestimmen.

$$\left[x_{lower\_limit}\,,\, x_{upper\_limit}\right] = \left(-\infty\,,\ \overline{x} + z_{0.95} \cdot \frac{s}{\sqrt{n}}\right]$$

Möglich ist es auch, den sogenannten **Verwerfungsbereich** anzugeben – das ist die Menge derjenigen Werte der Teststatistik, für die die Nullhypothese verworfen wird (siehe Abbildung 1). Die Wahrscheinlichkeit, unter der Nullhypothese einen Wert im Verwerfungsbereich zu erhalten, entspricht gerade dem Signifikanzniveau α. Das kann man etwa so sehen: Beträgt das Signifikanzniveau z.B. 5%, bedeutet das gerade, dass bei den 5% extremsten Werten („extrem“ unter der Nullhypothese) verworfen wird.

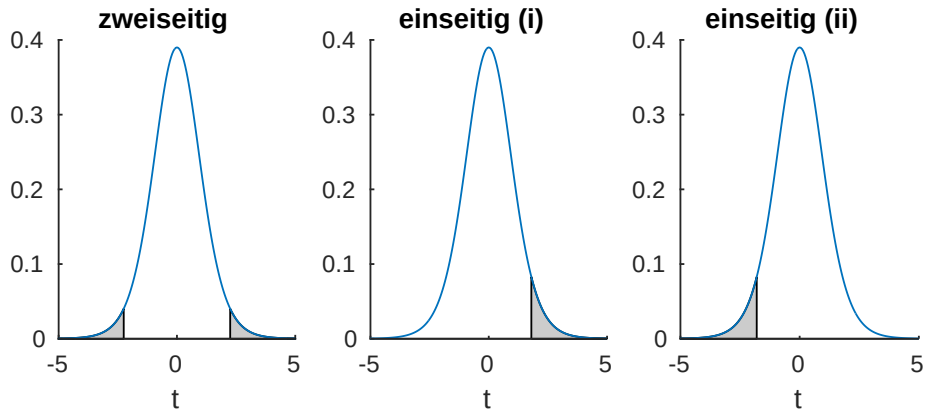

Abbildung 1 – Dichte der t-Verteilung mit 11 Freiheitsgraden. Die grauen Flächen machen jeweils 5% Wahrscheinlichkeit aus. Die entsprechenden Quantile auf der t-Achse ergeben gerade die Verwerfungsbereiche für den zwei- bzw. einseitigen t-Test auf dem Signifikanzniveau 5%.

7) **Testentscheid**: Falls der berechnete p-Wert *unter* dem Signifikanzniveau liegt, *verwerfen* wir die Nullhypothese (d.h., akzeptieren die Alternativhypothese als plausiblere Erklärung der Daten); falls der p-Wert *über* α liegt, *behalten* wir die Nullhypothese bei (d.h., wir halten die Daten für *nicht aussagekräftig* genug, um unsere Nullhypothese zu verwerfen – im Zweifel *für* den Angeklagten!).

   Im Fall der Abfüllmaschine ist der p-Wert von $p \approx 6{\cdot}10^{-12}$ viel kleiner als $\alpha = 0.05$, daher verwerfen wir die Nullhypothese, dass die Maschine korrekt arbeitet. Die Daten stützen die Vermutung des Müesli-Produzenten und die durchschnittliche Füllmenge ist (statistisch) *signifikant* (auf einem Signifikanzniveau von 5%) verschieden von 500 g.

Schritte 5 und 7 kann man für ein 5%-Signifikanzniveau konkret wie folgt zusammenfassen: Wenn wir unter der Nullhypothese in *weniger als* 5% *der Fälle* ein Ergebnis erwarten, das mindestens so extrem ist wie das gemessene, verwerfen wir die Nullhypothese; andernfalls behalten wir sie bei.

Will man also mit einem statistischen Test etwas „beweisen“, muss man die Nullhypothese so ansetzen, dass man sie (hoffentlich) verwerfen kann. Mit dem Signifikanzniveau können wir die Wahrscheinlichkeit kontrollieren, mit der das fälschlicherweise geschehen wird.

Der Einstichproben-t-Test

Der oben beschriebene z-Test hat zwei entscheidende Nachteile. Man schätzt die wahre unbekannte Varianz σ^2 mit Hilfe der Stichprobenvarianz s, also begeht man einen „hoffentlich“ kleinen Fehler. Und als zweites ist der z-Test nur sinnvoll, wenn der Stichprobenumfang n hinreichend gross ist (gemäss Faustregel ≥ 30), da nur dann die Annahme einer Normalverteilung des Mittelwerts und das Ersetzen von σ durch s (näherungsweise) gerechtfertigt ist.

In der Realität verwendet man aber fast immer den im Folgenden beschriebenen t-Test für den Mittelwert einer Stichprobe. Der t-Test funktioniert im Wesentlichen gleich wie der z-Test, unterscheidet sich aber dadurch, dass wir als Verteilung der Teststatistik erhält man nun nicht mehr die Standardnormalverteilung, sondern eine sogenannte (Student-)t-Verteilung (mit $n-1$ Freiheitsgraden).

Der Einstichproben-t-Test für kleinere Stichproben der Grösse ($10 \leq n \leq 30$)

Für eine kleine Stichprobe müssen wir nun auch eine Annahme über die Verteilung der einzelnen Messwerte treffen müssen (konkret nimmt man Normalverteilung an). Diese Annahme müssen wir immer Überprüfen.

Wir gehen die oben beschriebenen sechs Schritte nochmals durch und beschreiben die Änderungen im Detail. Wir bleiben beim Beispiel des Müesli-Produzenten, gehen nun aber von der folgenden Stichprobe der Grösse $n = 12$ aus:

Messung	1	2	3	4	5	6	7	8	9	10	11	12
Masse [g]	514	510	497	508	496	517	503	504	498	515	506	510

1) Wie bereits erwähnt benötigen wir für unser **Modell** nun die Zusatzannahme, dass die einzelnen Messungen normalverteilt sind. Formal: $X_1, X_2, \ldots, X_n$ sind i.i.d. *normalverteilt* mit Erwartungswert μ und Varianz σ^2 (beide unbekannt).

2) Unsere **Nullhypothese** ist weiterhin $H_0 : \mu = \mu_0 = 500$ (und die Alternativhypothese somit $H_A : \mu \neq \mu_0 = 500$).

3) Die **Teststatistik** ist weiterhin die standardisierte Grösse

$$T = \frac{\overline{X} - \mu_0}{S/\sqrt{n}} \ .$$

Die Überlegungen, weshalb dies grundsätzlich sinnvoll ist, sind die gleichen wie oben. Für die Stichprobe von $n = 12$ Messungen in unserem Beispiel erhalten wir einen Mittelwert von $\overline{x} = 506.5$ und eine empirische Standardabweichung von $s \approx 7.116$. Daraus resultiert ein Wert von $t \approx 3.164$ für die Teststatistik T.

Die Verteilung von T unter der Nullhypothese ist nun allerdings nicht mehr durch die Standardnormalverteilung, sondern durch die besagte *t-Verteilung* (mit $n-1$ Freiheitsgraden) gegeben. Es handelt sich dabei um eine um Null symmetrische Verteilung, welche von einem Parameter – der Anzahl *Freiheitsgrade* – abhängt. (Wir wollen uns an dieser Stelle nicht mit der genauen Definition der Dichtefunktion der t-Verteilung aufhalten.)

Wir erklären dafür nochmals kurz, warum T *nicht* standardnormalverteilt ist: Die Messungen sind ja gemäss angepasstem Modell i.i.d. normalverteilt mit Erwartungswert μ und Varianz σ^2. Aufgrund dieser Annahme ist $(\overline{X} - \mu_0)/(\sigma/\sqrt{n})$ (unter der Nullhypothese H_0) sogar exakt standardnormalverteilt. Da wir jedoch die Standardabweichung σ durch die empirische Varianz S schätzen müssen, und diese Schätzung bei kleinem n grosse Ungenauigkeit mit sich bringt, schlägt sich diese Ungenauigkeit auch in der tatsächlichen Verteilung unserer Teststatistik T nieder. Die t-Verteilung hat im Gegensatz zur Standardnormalverteilung mehr Gewicht auf den „Schwänzen“ – ist also „breiter“ (siehe Abbildung 2). Je grösser die Anzahl Freiheitsgrade, desto weniger Gewicht liegt auf den Schwänzen (da die Schätzung S für σ mit steigendem Stichprobenumfang präziser wird), und desto näher liegt die t-Verteilung an der Standardnormalverteilung. Für $n \to \infty$ strebt die t-Verteilung also gegen die Standardnormalverteilung, und wir erhalten wieder den oben bereits beschriebenen z-Test für grosse Stichprobenumfänge.

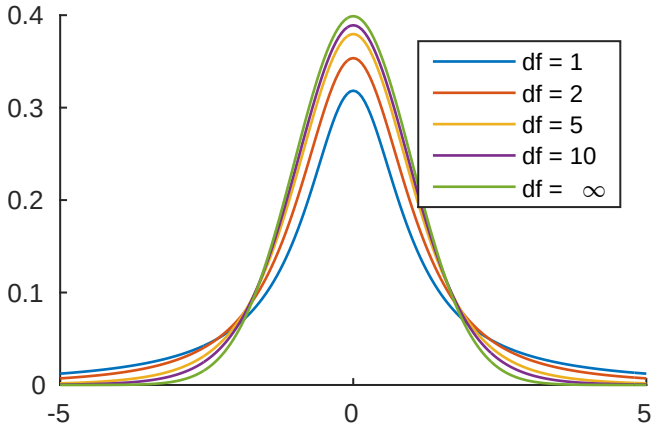

Abbildung 2 – Dichte der t-Verteilung für verschiedene Anzahl Freiheitsgrade („df“). Für df $= \infty$ entspricht die Verteilung gerade der Standardnormalverteilung.

4) Den zugehörigen **p-Wert** erhalten wir als

$$P(|T| \geq 3.164) = P(T \geq 3.164) + P(T \leq -3.164) = 2 \cdot (1 - P(T \leq 3.164)),$$

wobei T der t-Verteilung mit $n - 1 = 11$ Freiheitsgraden folgt.

Mit Python erhalten wir diese Wahrscheinlichkeit wie folgt:

```
> 2*(1 - stats.t.cdf(3.164, 11))
ans = 0.0090
```

5) Als **Signifikanzniveau** α wählen wir (in diesem Beispiel) weiterhin 5%.

6) Auch mit Hilfe der t-Verteilung können wir ein Konfidenzintervall für den Erwartungswert μ berechnen. Wir müssen hierzu einfach das Quantil der Standardnormalverteilung durch das entsprechenden Quantil der t-Verteilung ($t_{n-1,1-\alpha/2}$) ersetzen:

$$\left[\overline{x} - t_{n-1,1-\alpha/2} \cdot s/\sqrt{n}, \overline{x} + t_{n-1,1-\alpha/2} \cdot s/\sqrt{n}\right]$$

Für das konkrete Beispiel erhalten wir die Werte

```
> 506.5 - stats.t.ppf(0.025, df = 11)*7.116/sqrt(12)
ans = 501.9787
>   506.5 + stats.t.ppf(0.025, df = 11)*7.116/sqrt(12)
ans = 511.0213
```

Es ist nicht überraschend, dass dieses Konfidenzintervall (für die gleiche betrachtete Müeslimaschine!) deutlich breiter ist als im ersten Beispiel zum z-Test: Mit nur 12 Stichprobenwerten können wir mit der gleichen Sicherheit (95%) natürlich deutlich weniger präzise Aussagen machen als mit einer Stichprobengrösse von 50. Dementsprechend ist im zweiten Beispiel auch der p-Wert deutlich grösser als beim z-Test (aber immer noch klein genug, um die Nullhypothese zu verwerfen).

7) Der **Testentscheid** fällt somit gleich aus wie vorher: Es gilt immer noch $p < 0.05$, wir verwerden die Nullhypothese also, da wir (für das gewählte Signifikanzniveau) genügend statistische Evidenz dazu haben.

Sind Sie nun restlos überzeugt, dass die Müesli-Abfüllmaschine falsch kalibriert ist? Die gemachten Berechnungen basieren alle auf dem **Modell**, dass die Messungen normalverteilt sind. Diese Modellannahme müssen wir aber noch überprüfen, bevor wir einen endgültigen Entscheid treffen. Die Normalverteilung von Daten wird am besten graphisch mit einem sogenannten *Q-Q-Plot* überprüft. Hierbei werden die Datenpunkte in einem ersten Schritt der Grösse nach geordnet, welche nun gerade n empirischen Quantilen entsprechen. Als Vergleich dienen die entsprechenden theoretischen Quantile der Standardnormalverteilung. Die empirischen und theoretischen Quantile werden nun graphisch gegeneinander aufgezeichnet (siehe Abbildung 3).

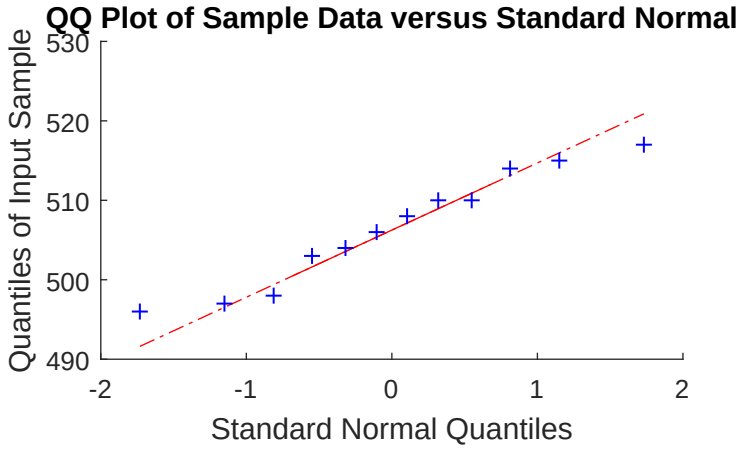

Abbildung 3 – Q-Q-Plot zur graphischen Überprüfung der Normalverteilung der Messungen der Müesli-Abfüllmaschine.

Wenn die Daten tatsächlich normalverteilt sind, existiert ein linearer Zusammenhang zwischen den empirischen und den theoretischen Quantilen ($Z = \mu + \sigma \cdot X$), d.h. die gezeichneten Punkte liegen auf einer Geraden. Grosse systematische Abweichungen von einer Geraden geben einen Hinweis darauf, dass die Daten nicht normalverteilt sind. Auf dem Cheat Sheet Hypothesentests finden Sie zudem weitere Tests mit denen wir die Normalität der Daten überprüfen können.

Betrachten wir nun Abbildung 3, so scheint die Annahme einer Normalverteilung plausibel. Dass hierbei der grösste und der kleinste Wert ein bisschen weiter weg von der Geraden liegen ist normal und nicht weiter tragisch. Wir sind nun also überzeugt, dass die Müesli-Abfüllmaschine (mit einer Sicherheit von mind. 95%) falsch kalibriert ist!

Bemerkungen:

In der Praxis müssen wir für viele Tests nicht von Hand die ganzen Berechnungen durchführen, sondern können vorgefertigte Funktionen von Python dafür verwenden. Der t-Test zum Beispiel ist in `scipy.stats` Package unter der Name `ttest` implementiert. Für den Datensatz aus dem Beispiel können wir den ganzen t-Test wie folgt durchführen:

```
> df = pd.DataFrame({"data": [514, 510, 497, 508, 496, 517,
   503, 504, 498, 515, 506, 510]})
> Teststatistik, pWert = stats.ttest_1samp(df, 500)
> print(Teststatistik)
ans = [3.1642641]
> print(pWert)
ans = [0.00901022]

# 95% Konfidenzinterval (erste Option)
> stats.t.interval(0.95, 49, loc=df.mean(), scale=df.std()/50**0.5)
ans = (array([504.47767699]), array([508.52232301]))

# 95% Konfidenzinterval (zweite Option)
> import statsmodels.stats.api as sms
> sms.DescrStatsW(df).tconfint_mean()
ans = (array([501.97875807]), array([511.02124193]))
```

Fehler 1. und 2. Art

Bei einem statistischen Test können wir zwei Arten von Fehlentscheiden treffen:

- Wir verwerfen die Nullhypothese, obwohl sie wahr wäre – das nennt man einen *Fehler 1. Art*. Die Wahrscheinlichkeit dafür kennt bzw. kontrolliert man. Sie entspricht genau dem gewählten Signifikanzniveau α.

- Wir behalten die Nullhypothese, obwohl sie falsch wäre – das nennt man einen *Fehler 2. Art*, bezeichnet mit β. Die Wahrscheinlichkeit dafür lässt sich kaum oder nur mit viel Aufwand bestimmen. Insbesondere muss man hierfür die wahre Verteilung der Messungen kennen, was in der Praxis fast nie der Fall ist.
  Die Gegenwahrscheinlichkeit des Fehlers 2. Art, also $1 - \beta$, entspricht der sogenannten *Power* eines statistischen Test. Die Power gibt uns an, wie hoch die

Wahrscheinlichkeit ist, dass wir mit einem statistischen Tests das Nichtzutreffen der Nullhypothese erkennen. Dieser Zusammenhang ist in Abbildung 4 visualisiert. Die Abbildung illustriert insbesondere auch den folgenden Zusammenhang: Je kleiner das Signifikanzniveau α, desto kleiner auch die Power $1 - \beta$.

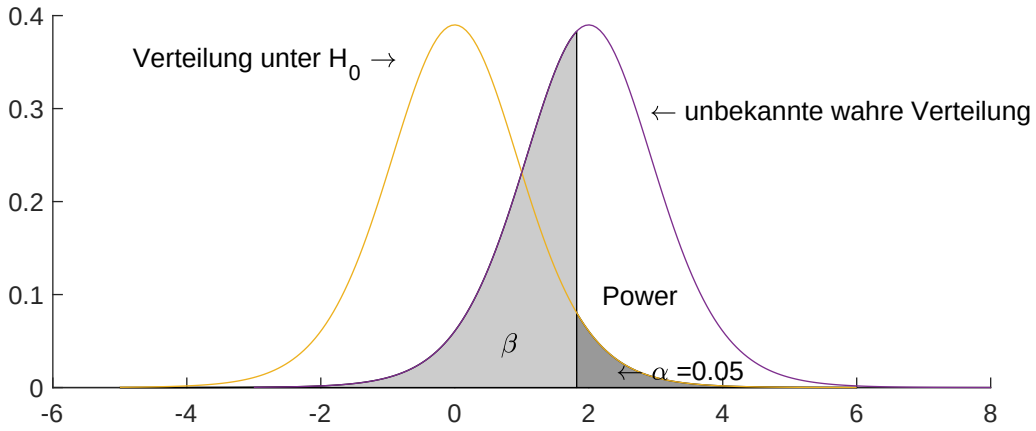

Abbildung 4 – Fehler 1. und 2. Art sowie Power eines einseitigen statistischen Tests.