Kapitel 3

Statistische Tests

3.1 Der α-Niveau-Test

3.1.1 Einführendes Beispiel

Ein Pharmaunternehmen hat 4000l einer teuren Reagenz bestellt, welche höchstens $7,5\%$ Verunreinigungen mit Wasser enthalten darf. Die Laboruntersuchung mit einem geeichten Verfahren der Abteilung für Qualitätskontrolle ergibt aber einen Wasseranteil von $x = 9.46\%$.

Natürlich unterliegt der Messwert X der Konzentration einer gewissen Messfehlerschwankung. Es sei (zur Vereinfachung für den Anfang) bekannt, dass der Messwert $P_\mu := N(\mu, 1\%^2)$ verteilt ist, wobei μ der wahre Wasseranteil ist.

Soll der Pharmakonzern die Lieferung reklamieren?

Problem: Bei normalverteilten Messwerten kann theoretisch bei jedem Wasseranteil jeder Messwert auftreten. Es kann also keinen definitiven Nachweis geben, dass die Wassermenge zu groß ist. Insbesondere werden wir auch nie genau $7,5\%$ messen.

Problemstellung

Es gibt zwei Möglichkeiten:

- **Hypothese:** Die Zulieferfirma liefert korrekt: $\mu = 7,5\%$
- **Alternative:** Die Zulieferfirma will sparen: $\mu > 7,5\%$

Bis zum Beweis des Gegenteils geht man davon aus, dass die Zulieferfirma korrekt liefert.

Entscheidungsfunktion: Wir stellen also eine Regel auf: Ist X in einem Annahmebereich $A \subset \mathbb{R}$, so nehmen wir die Lieferung an. Ist $X \notin A$, so lehnen wir die Lieferung ab. Das wird in eine Testfunktion kodiert:

$$T(X) = \begin{cases} 0, \text{ falls } X \in A \\ 1, \text{ falls } X \notin A \end{cases}$$

0 bedeutete annehmen, 1 bedeutet ablehnen.

Fehlerarten: Es gibt zwei Arten von Fehlentscheidungen:

	Lieferung korrekt	Lösung zu dünn
$T(X) = 0$	gute Lieferung wird angenommen richtige Entscheidung	schlechte Lieferung wird angenommen falsche Entscheidung Fehler 2.Art/β-Fehler
$T(X) = 1$	Annahme guter Lieferung verweigert falsche Entscheidung Fehler 1.Art/α-Fehler	Annahme mangelhafter Lieferung verweigert richtige Entscheidung

Regel: Beschränke den Fehler 1.Art:
Im Vertrag mit dem Zulieferer steht, dass die vom Pharmakonzern verwendete Prüfregel höchstens 5% der richtigen Lieferungen als falsch zurückschicken darf.
Regel: Minimiere den Fehler 2.Art:
Der Chef fordert, dass die Wahrscheinlichkeit für einen Fehler 2.Art, also für eine fälschlich akzeptierte Lieferung, so gering wie möglich sein sollte.

> Der Fehler erster Art wird also durch Festlegung beschränkt, der Fehler 2.Art nur durch Wünsche.

Festlegung des Annahmebereichs:
Je größer der Messwert, desto eher hat der Lieferant korrekt geliefert. Unser Annahmebereich hat wohl die Form $[k, \infty)$, dabei heißt k der kritische Wert (also der Wert an dem es sich entscheidet):

$$T(X) = \begin{cases} 0, \text{ falls } X > k \\ 1, \text{ falls } X \geq k \end{cases}$$

0 bedeutet, dass man die Hypothese akzeptiert und damit auch die Lieferung. 1 bedeutet, dass man die Hypothese widerlegt und daher die Lieferung zurückweist.

Die Lieferung wird also mit einer Wahrscheinlichkeit von

$$P_\mu((-\infty, k)) = F_{P_\mu}(k)$$

akzeptiert und mit einer Wahrscheinlichkeit von

$$P_\mu([k, -\infty)) = 1 - F_{P_\mu}(k)$$

abgelehnt.

Dabei soll (laut Chef) die Ablehnwahrscheinlichkeit so groß wie möglich sein, unter der Bedingung, dass (laut Vertrag) für $\mu = 7,5$ die Ablehnwahrscheinlichkeit nicht größer als $\alpha = 0.05$ sein darf.

Also gilt mit der Verteilungsfunktion F:

$$\begin{aligned} \alpha &= 1 - F_{P_{7,5}}(k) \\ F_{P_{7,5}}(k) &= 1 - \alpha \\ k &= F^{-1}_{P_{7,5}}(1 - \alpha) \end{aligned}$$

Also ist $k = F^{-1}_{P_{7,5}}(1-\alpha)$ zu wählen. Die Umkehrfunktion der Verteilungsfunktion ist die Quantilfunktion (hier die `qnorm` Funktion in R)

Der Wert für k ist also:

```
> qnorm(1 - 0.05, mean = 7.5, sd = 1)

[1] 9.145

> k = qnorm(1 - 0.05, mean = 7.5, sd = 1)
> pnorm(k, mean = 7.5, sd = 1)

[1] 0.95
```

Da der tatsächliche Messwert $8,45$ diesen Wert 9.145 nicht übersteigt, ist somit $T(x) = 0$ und die Lieferung wird akzeptiert.

3.1.2 Grundaufbau eines Tests

Wir haben den Grundaufbau eines Tests an einem vereinfachten Gauss-Test kennengelernt:

Zu jedem Test gehören eine Reihe von Elementen, die man kennen muss, um ihn erfolgreich anzuwenden:

- Die **Testsituation** beschreibt die grundsätzliche Situation für die ein Test gemacht ist.

  In unserem Beispiel: Feststellen, ob der wahre Mittelwert einer Gaussverteilten Zufallsgröße X oberhalb einer Grenze μ_0 liegt, wenn die Varianz σ_0^2 bekannt ist.

- Das **Testproblem** formuliert die genaue Fragestellung des Tests mittels eines Gegensatzpaares aus einer **Hypothese** (H_0) und einer **Alternative** (H_1).

  In unserem Beispiel:
  $H_0 : \mu \leq 7,5$ versus $H_1 : \mu > 7,5$

- Die **Voraussetzungen** des Tests formulieren das genaue statistische Modell, das gelten muss, damit der Test sich korrekt verhält.

  In unserem Beispiel:
  $X \sim N(\mu, \sigma_0^2)$

- Ein α**-Fehler**, der die Obergrenze für den Fehler 1.Art darstellt.

  In unserem Beispiel:
  $$\alpha = 0.05$$

  Dieses α-Niveau von $\alpha = 0.05$ ist weit verbreitet und werden wir als Standard-α-Niveau für unsere Vorlesung verwenden.

  Definition 30 *Ein Test, der, wenn die Hypothese gilt, diese mit einer Wahrscheinlichkeit kleiner/gleich α ablehnt, heißt ein α–Niveau–Test.*

- Eine **Entscheidungsregel**, die sagt, wie die Entscheidung des Tests aus den Daten berechnet werden kann. Der Test kann zwei Entscheidungen treffen:

  0 Der Test nimmt die Hypothese an.

  1 Der Test lehnt die Hypothese ab.

  Im einfachsten Fall besteht die Entscheidungsregel aus:

  – einer **Teststatistik**. Die Teststatistik ist eine Zufallsvariable, deren Realisierung aus den Daten berechnet werden kann.

  – einem **kritischen Wert**. Der Test „lehnt die Hypothese ab", falls die Teststatistik größer/gleich dem kritischen Wert ist.

  Der kritische Wert k kann in diesem Fall (bei stetiger Verteilungsfunktion) als das maximale $1 - \alpha$-Quantil der Verteilungen der Teststatistik unter der Hypothese berechnet werden, da dadurch der Fehler 1.Art auf α festgelegt wird. In unserem Fall (wie meistens) ist das $1 - \alpha$-Quantil maximal, wenn der Fall $\mu = 7,5$ als am Rand der Hypothese zu den Berechnungen herangezogen wird.

In unserem Beispiel: $T(X) = X$ und $k = F^{-1}_{N(\mu_0,\sigma_0^2)}(1-\alpha) = 9.1449$

```
> qnorm(1 - 0.05, mean = 7.5, sd = 1)

[1] 9.145
```

Im gegeben Beispiel ist also

$$9.46 = T(x) \geq k = 9.1449$$

und damit würde die Lieferung zurückgeschickt werden und die Annahme, der Zulieferer hätte korrekt geliefert, verworfen werden.

3.2 Weitere Grundbegriffe der Testtheorie

3.2.1 Interpretation als wissenschaftlicher Nachweis

- In den meisten Wissenschaftsbereichen gibt es eine akzeptierte Irrtumswahrscheinlichkeit α, die bestimmt welche Fehler 1.Art allgemein als erlaubt gelten. In den meisten Fällen (z.B. Biologie, Medizin, Geowissenschaften) ist $\alpha = 0.05 = 5\%$. In besonderes anspruchsvollen Bereichen (z.B. Sicherheitstechnik, Kerntechnik) werden nur kleinere α-Fehler z.B. $\alpha = 0.01$ oder $\alpha = 0.001$ verwendet. In der Datenerhebung besonders schwer zugänglichen Bereichen , z.B. in manchen Bereichen der Sozialpsychologie, wird mitunter auch $\alpha = 0.10 = 10\%$ verwendet.

- Wenn ein α-Nivau Test, die Hypothese ablehnt, so gilt dass als Nachweis, dass die Hypothese **nicht stimmt**, da ein Ablehnen sehr unwahrscheinlich ist, wenn die Hypothese gestimmt hätte.

Vorsicht: Diese Interpretation stimmt nur, wenn die Voraussetzungen des Tests tatsächlich erfüllt sind (Wüste der unerfüllten Voraussetzungen).

  Um zu klären, welches α verwendet wurde, spricht man dann z.B. vom „Nachweis auf dem 5%-Nivau“.

- Wenn ein α-Nivau Test, die Hypothese annimmt, so bedeutet das erst einmal gar nichts, da wir keinerlei Obergrenzen für den Fehler zweiter Art garantieren können. Eine typische Obergrenze ist $1 - \alpha \approx 1$.

Vorsicht: Wird der Test angenommen, so kann **nichts!!!** gefolgert werden. (Nacht der angenommenen Hypothesen).

  Gegebenfalls kann also das Wort “Nichts” die richtige Antwort auf eine Klausurfrage seine.

3.2.2 Gütefunktion

Haben wir unseren Test, dann wollen wir natürlich wissen, wie gut er ist, d.h. wie wahrscheinlich wir eine falsche Hypothese tatsächlich erkennen. Diese Wahrscheinlichkeit ist allerdings vom wahren Parameter abhängig.

Definition 31 *Dazu definieren wir die* **Gütefunktion**, *welche die Ablehnwahrscheinlichkeit in Abhängigkeit vom Parameter angibt:*

$G(\mu) = E_{P_\mu}[T(X)] =$ *Wahrscheinlichkeit, dass der Test ablehnt, wenn* μ *der richtige Parameter ist.*

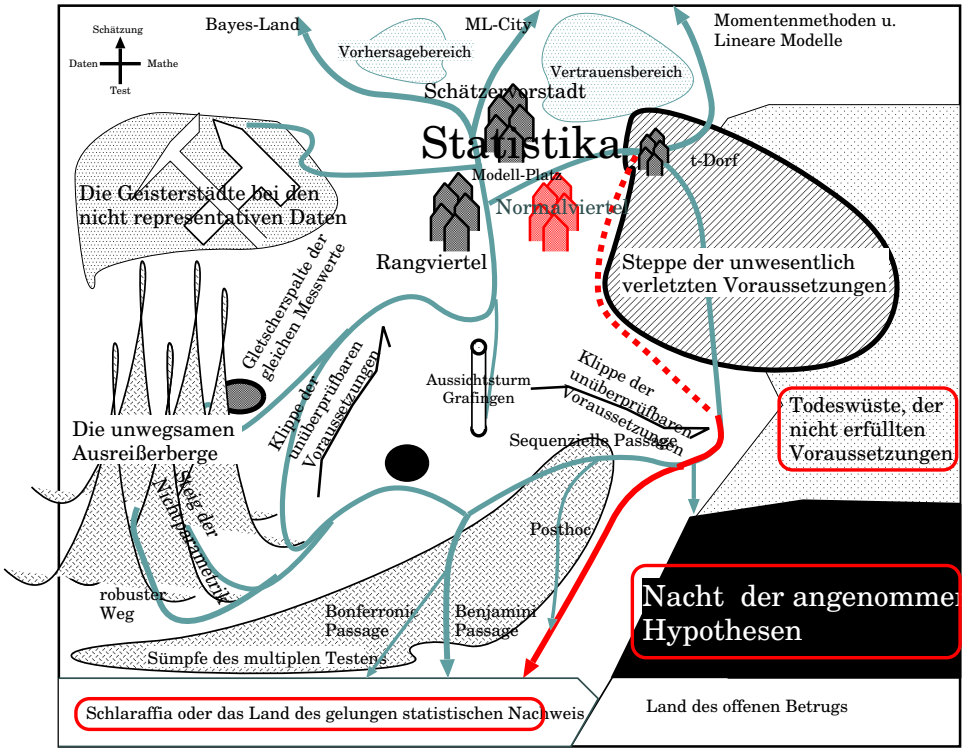

Abbildung 3.1: Übersichtskarte Grundbegriffe der Testtheorie
Die Einordnung von Kapitel 3.2 in die Übersichtskarte.

An der Gütefunktion kann man erkenne, dass die Wahrscheinlichkeit für einen Fehler 2.Art, also die Wahrscheinlichkeit für ein fälschliches Annehmen der Hypothese sehr hoch sein kann. Je höher die Gütefunktion ist, desto wahrscheinlicher gelingt der Nachweis, desto wahrscheinlicher kann man also etwas nachweisen. Der Wert der Gütefunktion für einen speziellen Parameter heißt daher auch **Macht** oder **Power**.

Definition 32 *Ein Test für den die Ablehnwahrscheinlichkeit $G(\mu)$ für jeden Parameterwert μ aus der Alternative $> \alpha$ ist, heißt* **unverzerrt** *(***unbiased***) .*

3.2.3 Die Sümpfe des multiplen Testens

Herr Streber ist neu in der Qualitätssicherung und möchte die Eingangskontrolle verbessern. Dazu schlägt er vor, mit entsprechenden Testverfahren, nicht nur die Grenzwerte für den Wasseranteil, sondern für alle 200 Chemikalien und Spurenelemente, die im Labor gemessen werden können, die Anteile zu überprüfen und im Falle des Nachweises einer Erhöhung auf dem 5%-Niveau die Lieferung anzulehnen.

Wie wahrscheinlich ist es nun, dass die Lieferung zurückgeschickt wird?

Nehmen wir einmal an, der Zulieferer hat korrekt (und immer genau am Grenzwert) geliefert. Dann ist für jeden gemessenen Stoff die Wahrscheinlichkeit, dass ein hoher Wert, der zur Ablehnung der Lieferung für genau 5%:

$$P(T_i(X) = 1) = 0.05 = \alpha \text{ für } i = 1, \ldots, 200$$

und somit

$$P(T_i(X) = 0) = 0.95 = 1 - \alpha \text{ für } i = 1, \ldots, 200$$

```
> G <- function(mu) 1 - pnorm(k, mean = mu, sd = 1)
> mu <- seq(0, 15, by = 0.01)
> plot(mu, G(mu), xlab = expression(mu), ylab = expression(G(mu)),
+     type = "n", main = "Guetefunktion unseres Tests",
+     lwd = 3)
> segments(0, 1.01, 7.5, 1.01, col = "green", lwd = 3)
> segments(7.5, 1.01, 15, 1.01, col = "red", lwd = 3)
> abline(v = 7.5)
> abline(h = 0.05)
> abline(h = c(0, 1), col = "gray")
> lines(mu, G(mu), lwd = 3)
> text(7.5/2, 1, expression("Hypothese " * H[0]), pos = 1,
+     col = "green")
> text((7.5 + 15)/2, 1, expression("Alternative " * H[1]),
+     pos = 1, col = "red")
> text(2, 0.05, expression(alpha == 0.05), pos = 3)
> text(7.5, 0.5, expression(mu[0] == 7.5), pos = 2)
```

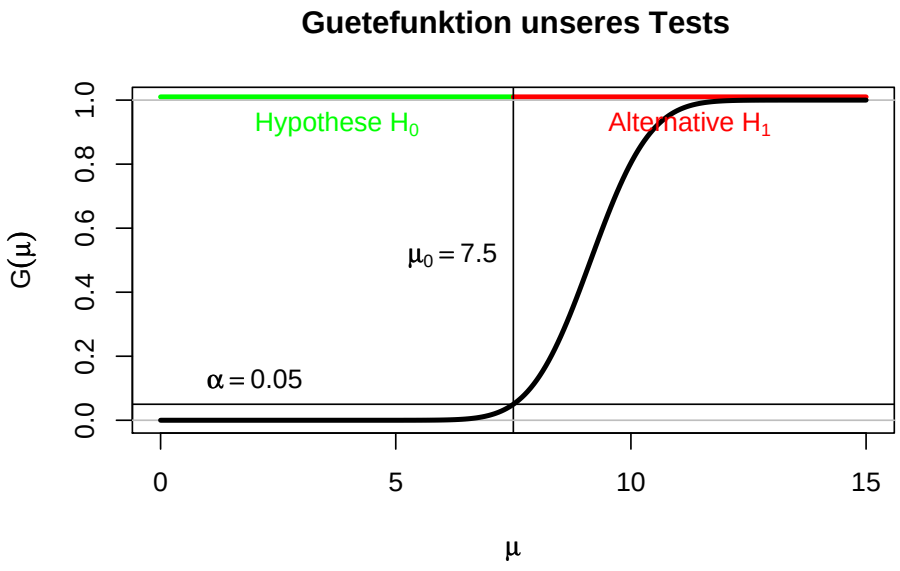

Abbildung 3.2: Diese Graphik zeigt die Gütefunktion des vereinfachten Gauss-Test. Die Gütefunktion entspricht der Wahrscheinlichkeit, bei einem gegebenen Parameter die Hypothese abzulehnen.
Die vertikale Linie bezeichnet die Grenzlinie der Hypothese und der Alternative. Die horizontale Linie bezeichnet das α-Niveau von 5%. Im Bereich der Hypothese liegt die Gütefunktion unter dem α-Niveau. An der Grenze der Hypothese liegt sie oft genau auf α. Im Bereich der Alternative – rechts – möchte man möglichst große Werte erreichen, da man ja eine falsche Hypothese mit hoher Wahrscheinlichkeit ablehnen möchte. Im Übergangsbereich ist diese Wahrscheinlichkeit allerdings eher gering. Erst für deutlich zu hohe Werte wird auch die Ablehnwahrscheinlichkeit groß und nähert sich dann schnell 1 an. Eine Test dessen Gütefunktion im Bereich der Alternative immer über α liegt, heißt unverzerrt.

Da die Labormessungen unabhängig stattfinden, kann man davon ausgehen, dass die Ablehnung unabhängig erfolgt. Die Wahrscheinlichkeit dafür, das alle Test angenommen werden ist also:

$$P(T_i(X) = 0, i = 1, \ldots, 200) = \prod_{i=1} P(T_i(X) = 0) = (1-\alpha)^2 00 = 0.95^{200} = 3.50526662488287e{-}05$$

Die Lieferung würde also mit an Sicherheit grenzender Wahrscheinlichkeit abgelehnt werden.

Man würde sogar erwarten, dass bei dieser Vorgehensweise im Durchschnitt

$$E\left[\sum_{i=1}^{200} T_i(X)\right] = \sum_{i=1}^{200} E\left[T_i(X)\right] = 200 * 0.05 = 10$$

Verunreinigungen nachgewiesen werden.

Das Vorgehen von Herrn Streber führt also weder zu einem sinnvollen Eingangskontrollverfahren noch, selbst wenn bekannt ist, dass eine Verunreinigung vorliegt, zu einer gesicherten Information, durch welchen Stoff sie erfolgt.

Werden mehrere Tests durchgeführt, so erhöht sich die Wahrscheinlichkeit falsch signifikanter Ergebnisse.

> Ohne weitere Korrekturmaßnahmen darf zu jeder wissenschaftlichen Gesamtfragestellung also nur jeweils ein α-Nivau-Test durgeführt werden. Dieses Problem nennt man das Problem des **multiplen Testens**.

Eine Reihe von Auswegen werden wir in Abschnitt 3.6 besprechen.

3.2.4 Der p-Wert eines Tests

Der Computer weiß normalerweise nicht, welches α-Niveau Sie anstreben. Daher wird auch nicht ausgegeben, ob der Test annimmt oder ablehnt. Statt dessen wird ausgegeben welches das kleinste α-Niveau ist, zu dem der Test gerade noch ablehnt. Der ausgegebene Wert wird als p-**Wert** (**p-value**) bezeichnet.

> Die Interpretation erfolgt also nach dem Schema:
>
> $$p \leq \alpha \Leftrightarrow \text{Hypothese wird vom Test abgelehnt.}$$
>
> Wer das nicht weiß, fällt durch die Klausur!!!

Eine R-Funktion für so einen Test würde also so aussehen:

```
> EinfacherGauss.test <- function(x, mean = 0, var = 1) {
+     parameter <- c(mean = mean, sd = sqrt(var))
+     statistic <- c(T = x)
+     structure(list(data.name = deparse(substitute(x)),
+         method = "Ein Stichproben Gauss-Test", alternative = "greater",
+         parameter = parameter, statistic = statistic,
+         p.value = 1 - pnorm(statistic, mean = parameter["mean"],
+             sd = parameter["sd"])), class = "htest")
+ }
```

Vorgehensweise bei der Interpretation:

1. **Test heraussuchen**
   Für diesen Schritt müssen Sie verschiedene Tests mit ihren Anwendungssituationen kennen.

2. **Voraussetzungen prüfen:**
   Das Prüfen der Voraussetzungen kann z.B. die Durchführung weiterer Tests,

das Ansehen von Graphiken und das Nachlesen von Versuchsbeschreibung und anderen Quellen beinhalten.

In diesem Fall sind die Voraussetzungen des Tests, nach der Beschreibung des Beispiels, erfüllt.
Für diesen Schritt müssen Sie die Voraussetzungen des Tests kennen.

3. **Test durchführen:**

```
> EinfacherGauss.test(9.46, mean = 7.5, var = 1)

        Ein Stichproben Gauss-Test

data:  9.46
T = 9.46, mean = 7.5, sd = 1.0, p-value = 0.025
alternative hypothesis: greater
```

Für diesen Schritt müssen Sie mit einer Statistik-Software vertraut sein.

4. **p-Wert interpretieren:**
$0.025 \leq 0.05$ ist, lehnt der Test die Hypothese auf dem 0.05-Niveau ab.
Für diesen Schritt müssen Sie $p \leq \alpha \Leftrightarrow$ abgelehnt unbedingt im Kopf haben!!!

5. **Formale Schlüsse ziehen:**
Da die Voraussetzungen erfüllt waren und die Hypothese abgelehnt wurde, wurde nachgewiesen, dass nicht $\leq 7.5\%$-Wasser in der Lösung sind.
Für diesen Schritt müssen Sie die Hypothese des Tests kennen.

6. **Inhaltliche Schlüsse ziehen:**
Es wurde also nachgewiesen, dass zu viel Wasser in der Lieferung ist.
Für diesen Schritt müssen Sie den Anwendungskontext verstehen.

Details: Für einen Test der Form (Test mit Teststatistik $S(X)$)

$$T(X) = \begin{cases} 0 & S(X) \leq c \\ 1 & S(X) > c \end{cases}$$

entspricht der p-Wert

$$p = P^X(\{X : S(X) > S(x)\}) = P^S((S(x), \infty)) = 1 - F_S(S(x))$$

wobei P die Wahrscheinlichkeitsverteilung unter der Hypothese darstellt. Ist die Wahrscheinlichkeitsverteilung unter der Hypothese nicht eindeutig, so gilt:

$$p = \sup_{P \in H_0} P^X(\{X : S(X) > S(x)\}) = \sup_{P \in H_0} P^S((S(x), \infty)) = \sup_{P \in H_0} 1 - F_S(S(x))$$

3.2.5 Bezeichnungen

Hinter statistischen Tests verbirgt sich eine etwas verdrehte Logik mit einigen Verneinungen:

> Angenommen die Voraussetzungen des Tests sind erfüllt, dann gilt: "Ist die Beobachtung unter der Voraussetzung, dass die Hypothese stimmt, sehr unwahrscheinlich $p < \alpha$, so kann die Hypothese nicht gestimmt haben und ist somit widerlegt."

Um trotzdem klar formulieren zu können, haben sich eine Reihe verschieden aufwendiger Formulierungen eingebürgert, die verwendet werden, um die eine oder andere Situation zu beschreiben:

Wenn wir einen 0.05-Niveau Test verwendet haben, und $p \leq 0.05$ herausgekommen ist, so sagt man:

- Der Test hat die Hypothese auf dem α-Niveau von 5% signifikant abgelehnt.
- *Der Test hat die Hypothese auf dem 5%-Niveau signifikant abgelehnt.
- Der Test hat die Hypothese signifikant abgelehnt.
- Der Test hat die Hypothese abgelehnt.
- Der Test hat abgelehnt.
- *Der Test ist signifikant.
- Die Hypothese wurde auf dem α-Niveau von 5% abgelehnt.
- *Die Hypothese wurde auf dem 5%-Niveau abgelehnt.
- *Die Hypothese wurde abgelehnt.
- Der Test hat sich für die Alternative entschieden.
- Der Test hat sich (auf dem 5%)-Niveau signifikant für die Alternative entschieden.
- Der Test war auf dem 5%-Niveau signifikant.
- Wir haben Signifikanz.
- Es war signifikant.
- Hypothese abgelehnt.

(Die mit * markierten Texte sind nach meinem Stilempfinden die im normalen Gebrauch sinnvollsten Formulierungen.)

Ist jedoch herausgekommen $p > 0.05$, so sagt man:

- Der Test hat die Hypothese bei einem α-Niveau von 5% angenommen.
- *Der Test hat die Hypothese auf dem 5%-Niveau angenommen.
- Der Test hat die Hypothese angenommen.
- *Die Hypothese wurde angenommen.
- Der Test hat sich für die Hypothese entschieden.
- Der Test war auf dem 5%-Niveau nicht signifikant.
- *Der Test war nicht signifikant.

In keinem Fall sind die folgenden daraus zusammengesetzten, aber sinnentstellenden Phrasen zu verwenden:

- Die Alternative wurde angenommen.
  Diese Formulierung suggeriert dass die Beobachtung mit der Alternative kompatibel ist, was aber gar nicht geprüft wurde.

- Die Alternative wurde abgelehnt.
  Diese Formulierung suggeriert einen irgendwie gearteten Nachweis, dass die Alternative widerlegt sei.
- Die Hypothese ist ... signifikant.
  Unklar bleibt: angenommen oder abgelehnt. Nur der Test kann signifikant sein.
- Die Alternative ist ... signifikant.
  Vermutlich meint da jemand: Die Hypothese wurde signifikant abgelehnt und der Test hat sich signifikant für die Alternative entschieden. Die Formulierung suggeriert jedoch, dass dadurch die Alternative statistisch nachgewiesen sei.
- Das α-Nivau ist signifikant.
  Purer Unsinn. Findet sich aber leider oft in Klausurantworten.
- Der Test wurde abgelehnt.
  Das bedeutet eigentlich, dass jemand der Meinung ist, dass der Test für die gegebenen Situation nicht geeignet ist.

3.2.6 Testen auf einer Stichprobe: Der Ein-Stichproben-Gausstest

Ein erweitertes Testproblem ergibt sich, wenn der Pharmakonzern beschließt, dass der Wassergehalt zu genaueren Bestimmung mehrfach gemessen werden soll:

- **Einseitiger Ein-Stichproben-Gausstest**

  Situation: Test auf Mittelwert bei bekannter Varianz

  $H_0 : \mu = \mu_0$

  $H_1 : \mu > \mu_0$

  Voraussetzungen: $X_i \sim N(\mu, \sigma_0^2)$

  Bemerkung: Der Gauss-Test wird sehr selten auf reale Datensätze angewendet, da die Varianz fast nie bekannt ist. Er ist jedoch der wohl am leichtesten theoretisch zu verstehende Test und daher immer noch überall zu finden.

Um eine ganze Stichprobe in den Test mit einzubeziehen, verwendet man eine Statistik $S(X)$ deren Verteilung sich zwischen Hypothese und Alternative möglichst gut unterscheidet, anstelle des einzelnen Datenwertes:

$$T(X) = \begin{cases} 0 & S(X) \leq c \\ 1 & S(X) > c \end{cases}$$

S heißt dann die Teststatistik. Für den Ein–Stichproben–Gausstest verwendet man beispielsweise den Mittelwert der Stichprobe:

$$S(X) = \frac{1}{n} \sum_{i=1}^{n} X_i$$

der wenn X_i $N(\mu, \sigma^2)$ i.i.d., dann $N(\mu, \frac{1}{n}\sigma^2)$ verteilt ist. Diesen Wert kann man dann sozusagen als einzelne Beobachtung in den einfachen Gausstest einsetzten:

$$T(X) = \begin{cases} 0 & \frac{1}{n} \sum_{i=1}^{n} X_i \leq c \\ 1 & \frac{1}{n} \sum_{i=1}^{n} X_i > c \end{cases}$$

wobei der kritische Wert wie für eine einzelne "Beobachtung" mit Varianz $\frac{1}{n}\sigma^2$ berechnet wird.

```
> EinStichprobenGauss.test <- function(x, mean = 0, var = 1) {
+     xQuer <- mean(x)
+     n <- length(x)
+     parameter <- c(sd = sqrt(var/n))
+     statistic <- c(S = (xQuer - mean))
+     structure(list(data.name = deparse(substitute(x)),
+         method = "Ein Stichproben Gauss-Test", alternative = "greater",
+         parameter = parameter, estimate = xQuer, statistic = statistic,
+         p.value = 1 - pnorm(statistic, sd = parameter["sd"])),
+         class = "htest")
+ }
> EinStichprobenGauss.test(rnorm(20, mean = 10), mean = 7.5,
+     var = 1)

        Ein Stichproben Gauss-Test

data:  rnorm(20, mean = 10)
S = 2.411, sd = 0.224, p-value < 2.2e-16
alternative hypothesis: greater
sample estimates:
[1] 9.911
```

3.2.7 Alternative Alternativen

Es gibt drei verschiedene Arten von Alternativen, wenn auf einen speziellen Mittelwert getestet werden soll:

- "größer":
  $H_0 : \mu = \mu_0$ vs. $H_1 : \mu > \mu_0$
  Wie in unserem Beispiel mit $\mu_0 = 7,5$.

$$T(X) = \begin{cases} 0 & \frac{1}{n}\sum_{i=1}^{n} X_i \leq c \\ 1 & \frac{1}{n}\sum_{i=1}^{n} X_i > c \end{cases}, c = F^{-1}_{N(\mu, \frac{1}{n}\sigma^2)}(1-\alpha)$$

- "kleiner":
  $H_0 : \mu = \mu_0$ vs. $H_1 : \mu < \mu_0$
  Wenn das Unterschreiten eines Grenzwertes μ_0 nachgewiesen werden soll.

$$T(X) = \begin{cases} 0 & \frac{1}{n}\sum_{i=1}^{n} X_i \geq c \\ 1 & \frac{1}{n}\sum_{i=1}^{n} X_i < c \end{cases}, c = F^{-1}_{N(\mu, \frac{1}{n}\sigma^2)}(\alpha)$$

- "ungleiche" oder "zweiseitig"
  $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \neq \mu_0$
  Wenn gezeigt werden soll, dass der wahre Mittelwert nicht μ_0 entspricht.

$$T(X) = \begin{cases} 0 & \text{sonst} \\ 1 & \frac{1}{n}\sum_{i=1}^{n} X_i > c_o \\ 1 & \frac{1}{n}\sum_{i=1}^{n} X_i < c_u \end{cases}, c_u = F^{-1}_{N(\mu, \frac{1}{n}\sigma^2)}(\frac{1}{2}\alpha), c_o = F^{-1}_{N(\mu, \frac{1}{n}\sigma^2)}(1-\frac{1}{2}\alpha)$$

```
> Gauss.test <- function(x, mean = 0, var = 1, alternative = c("two.sided",
+     "less", "greater")) {
+     xQuer <- base::mean(x)
+     n <- length(x)
```

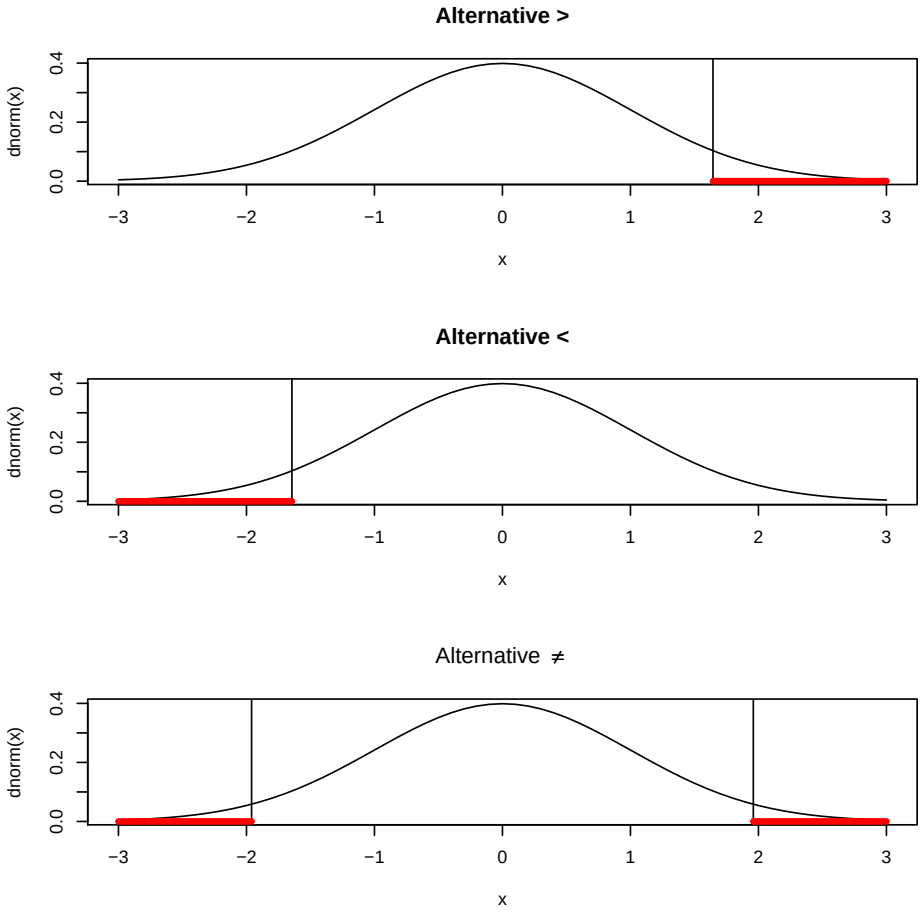

Abbildung 3.3: Ablehungsbereiche der verschiedenen Alternativen beim Gauss-Test. Der Ablehnbereich hat jeweils eine Wahrscheinlichkeit von α unter der Nullhypothese.

```
+     parameter <- c(sd = sqrt(var/n))
+     statistic <- c(S = (xQuer - mean))
+     alternative <- match.arg(alternative)
+     if (alternative == "greater")
+         p.value <- 1 - pnorm(statistic, sd = parameter["sd"])
+     else if (alternative == "less")
+         p.value <- pnorm(statistic, sd = parameter["sd"])
+     else if (alternative == "two.sided")
+         p.value <- 2 * min(1 - pnorm(statistic, sd = parameter["sd"]),
+             pnorm(statistic, sd = parameter["sd"]))
+     else stop("Unbekannter Alternativentyp")
+     structure(list(data.name = deparse(substitute(x)),
+         method = "Ein Stichproben Gauss-Test", alternative = alternative,
+         parameter = parameter, estimate = xQuer, statistic = statistic,
+         p.value = p.value), class = "htest")
+ }
> Gauss.test(rnorm(20, mean = 10), mean = 7.5, var = 1,
+     alternative = "greater")

        Ein Stichproben Gauss-Test

data:  rnorm(20, mean = 10)
S = 2.448, sd = 0.224, p-value < 2.2e-16
alternative hypothesis: greater
sample estimates:
[1] 9.948

> Gauss.test(rnorm(20, mean = 10), mean = 7.5, var = 1,
+     alternative = "less")

        Ein Stichproben Gauss-Test

data:  rnorm(20, mean = 10)
S = 2.669, sd = 0.224, p-value = 1
alternative hypothesis: less
sample estimates:
[1] 10.17

> Gauss.test(rnorm(20, mean = 10), mean = 7.5, var = 1)

        Ein Stichproben Gauss-Test

data:  rnorm(20, mean = 10)
S = 2.214, sd = 0.224, p-value < 2.2e-16
alternative hypothesis: two.sided
sample estimates:
[1] 9.714
```

Grundsätzlich gilt wieder der Grundsatz des multiplen Testens: Es darf (ohne Korrekturen) nur einer der Tests durchgeführt werden. Würde man z.B. auf "größer" und auf "kleiner" testen, so erhält man insgesamt eine Wahrscheinlichkeit von $min(2\alpha, 1)$ die Hypothese fälschlicherweise abzulehnen. Man muss sich also vor dem Test entscheiden, welche Alternativen interessant sind, d.h. in welchen Fällen man einen wahrscheinlichen Nachweis haben möchte.

Beispiel 33 *Für unseren Pharmakonzern ist eine Ablehnung der Lieferung im Fall eines Wasseranteils deutlich unter dem Grenzwert nicht erstrebenswert. Der Pharmakonzern macht also einen einseitigen Test auf "größer".*

3.2.8 Was man zu einem Test wissen muss

- **Name des Tests**

  Situation: Eine Beschreibung der Situation welcher der Test eingesetzt werden kann. Diese Beschreibung sollten Sie kennen, um in jeder Situation einen geeigneten Kandidaten für einen Test zu haben.

  H_0 : Das was man wiederlegen möchte.

  H_1 : Die Situation für die der Test gemacht ist.

  Voraussetzungen: Die Voraussetzungen, die Sie zur Anwendung des Tests überprüfen müssen

  Bemerkung: Was man sonst noch so über den Test wissen sollte.

Eine systematische Auflistung der wichtigsten Tests finden Sie im Anhang B.

3.2.9 Überblick über die Testsituationen

Die Testsituationen werden nach mehrere Kriterien unterteilt:

- **Anzahl der beteiligten Stichproben**
  Es gibt Test für
  - *Ein-Stichproben-Tests*: Es werden Eigenschaften einer Grundgesamtheit untersucht, z.B. der besprochene Gausstest.
  - *Zwei-Stichproben-Tests*: Es werden zwei Grundgesamtheiten verglichen, z.B. der Zwei Stichproben t-Test zum Vergleich der Erwartungswerte zweier Grundgesamtheiten.
  - *Mehr-Stichproben-Tests*: Es wird überprüft, ob mehrere Stichproben Grundgesamtheiten mit gleichen Eigenschaften entstammen, z.B. Bartlett-Test zum Vergleich der Varianzen mehrere Stichproben.
- **Zu testende Größe**
  - *Mittelwert/Lage*:
  - *Varianz/Streuung*:
  - *Verteilung*:
  - *Unabhängigkeit*:
- **Art der Alternative**
  - "Ungleich"
  - "Größer"
  - "Kleiner"

  Die Tests auf größer und kleiner heißen auch **einseitige Tests**, da von der Hypothese aus die Alternative nur in einer Richtung liegt. Tests bei denen die Alternative in beiden Richtungen von der Hypothese liegt heißen auch **zweiseitige Tests.**
- **Art der Voraussetzungen**

  - *Normalverteilung*
    Normalverteilungsbasierte Tests setzten die Normalverteilung der interessierenden Größen voraus. Sie sind meist einfach zu berechnen und funktionieren gut, solange die Verteilung ungefähr einer Normalverteilung ähnelt. Insbesondere sollte sie eingipflig sein, keine schweren Schwänze haben und es sollten keine Ausreißer vorliegen. Der Einsatz dieser Tests, wenn keine exakte Normalverteilung vorliegt, wird in Figur 3.2 als "die Steppe der unwesentlich verletzten Voraussetzungen" bezeichnet.
  - *Nichtparametrische Test/Rangtests*
    Rangtests basieren auf der Ersetzung der Daten durch Rangzahlen und benötigen keine exakten Verteilungsvoraussetzungen und haben meist eine geringere Power als die entsprechenden normalverteilungsbasierten Tests. Sofern möglich wird also eher der normalverteilungsbasierte Test eingesetzt. Da die Vergabe von Rangziffern bei mehrfachen gleichen Messwerten schwierig ist, werden in diesem Fall die berechneten p-Werte ungenau und die Anwendung dieser Tests problematisch. Verschiedene Softwarepakete können das Problem verschieden gut kompensieren. Man weiß allerdings selten welches Paket wie gut ist.

    Der Einsatz dieser Tests wird in Figur 3.2 als "die Steige der Nichtparametrik" bezeichnet. Das Problem der gleichen Messwerte ist die "Falle der gleichen Messwerte" in die man fällt, wenn man diese Methoden unbedacht anwendet.
  - *Robuste Tests*
    Robuste Tests stehen zwischen diesen beiden Extremen und setzten eine Normalverteilung voraus, die allerdings durch Ausreißer und schwere Verteilungsschwänze kontaminiert sein darf. Von der Power her stehen die Tests zwischen beiden Kategorien. Der Einsatz dieser Tests wird in Figur 3.2 als "der robuste Pfad" bezeichnet.

    Für den Anwender sind diese Tests oft noch schwierig zu handhaben. Es gibt noch nicht zu allen Situationen entsprechende kanonische robuste Tests und wenn es sie gibt, sind sie in der Software oft noch nicht verfügbar. Außerdem muss der Anwender meist den möglichen Anteil der Ausreißer selbst festlegen. Es steht jedoch zu erwarten, dass sich diese Situation in den nächsten Jahren deutlich bessern wird.

    Die Ergebnisse beziehen sich dann jeweils auf eine ausreißerfreie Grundgesamtheit.

- **Anzahl der beteiligten Merkmale**
  - *univariat*: **Es wird nur ein Merkmal betrachtet.**
  - *bivariat*: **Es werden zwei Merkmale betrachtet.**
  - *multivariat*: **Es werden mehrere Merkmale betrachtet.**

Wir werden in dieser Vorlesung nur univariate und einige wenige bivariate Methoden besprechen. Die multivariaten Methoden werden in der multivariaten Statistik behandelt.

Für die meisten Kombinationen dieser Situationen gibt es einen Test. Die Tests selbst heißen meist nach der Verteilung der Teststatistik oder nach ihrem Erfinder. Da die Verteilungen vieler Teststatistiken gleich sind, haben viele Tests Namen mit vielen Attributen z.B. Zwei-Stichproben-t-Test oder χ^2-Test auf Unabhängigkeit.

3.3 Die t-Tests und ihre Verwandten

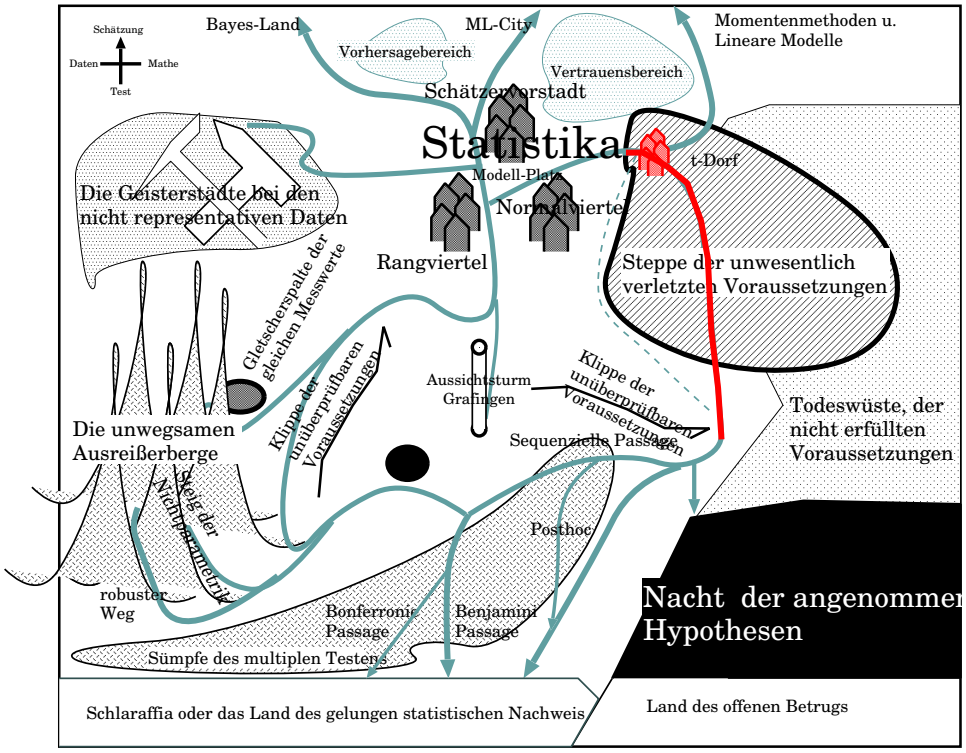

Abbildung 3.4: Übersichtskarte normalverteilungsbasierte Tests
Die Einordnung von Kapitel 3.3 in die Übersichtskarte.

3.3.1 Der Ein-Stichproben-t-Test

In praktischen Anwendungsfällen ist die Varianz nie bekannt. Unser Testproblem lautet also eher

$$H_0 : \mu = \mu_0 \text{ vs. } H_1 : \mu > \mu_0$$

unter der Voraussetzung $X_i \sim N(\mu, \sigma^2)$ für irgendein $\sigma^2 > 0$.

Ist aber die Varianz σ^2 unbekannt, so ist auch die Verteilung des Mittelwertes unbekannt. Ein solcher zusätzlicher für die eigentliche Fragestellung nebensächlicher Parameter heißt **Nebenparameter** (**nuisance** parameter, = "störender Parameter").

Die von William Sealey Gosset (1876-1937) unter dem Pseudonym "Student" publizierte Lösungsidee ist einfach und auch in vielen weiteren Fällen anwendbar: Man wähle eine Statistik, deren Verteilung unter Änderung des Nebenparameters invariant bleibt. Solche eine Statistik heißt **Pivot–Statistik**.

In diesem Fall kann man dass erreichen, indem man die Abweichung vom Vergleichswert durch die für sie geschätzte Varianz teilt:

$$S(X) := \frac{\bar{X} - \mu_0}{\sqrt{\frac{1}{n}\hat{\sigma}^2}}, \text{ mit } \hat{\sigma}^2 := \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Die Unabhängigkeit der Verteilung dieser Statistik unter der Nullhypothese $E[X] = \mu_0$ von σ^2 kann man erkennen, wenn man X durch ein Y mit einer größeren Streuung ersetzt: $Y := s(X - \mu_0) + \mu_0$

$$\hat{\sigma_Y^2} = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = s^2\hat{\sigma^2}$$

$$S(Y) = \frac{s(X - \mu_0) + \mu_0 - \mu_0}{\sqrt{s^2\frac{1}{n}\hat{\sigma}^2}} = S(X)$$

Durch die zufällige Normalisierung mit $\hat{\sigma}^2$ wird jedoch eine zusätzliche Variabilität erzeugt. Die Statistik ist also nicht mehr normalverteilt. Insbesondere werden in dem Fall, in dem die Varianz stark unterschätzt wird, eventuell sehr extreme Werte beobachtet.

3.3.2 Relevante Verteilungen

Wir müssen uns also mit Verteilungen des Varianzschätzers und Quotienten dieser mit normalverteilten Größen auseinandersetzten.

Definition 34 *Seien $X_1, \ldots, X_d$ und $Y_1, \ldots, Y_m$ unabhängig identisch $N(0,1)$-verteilt. So heißt die Verteilung von*

- $S_X^2 = \sum_{i=1}^{n} X_i^2$*, eine* **$\chi^2$-Verteilung** *mit d Freiheitsgraden (sprich: Chi-Quadrat-Verteilung) (in Zeichen χ_d^2)*
- $\frac{Y_1}{\sqrt{\frac{1}{n}S_X^2}}$*, eine* **(Student-)t-Verteilung** *mit d Freiheitsgraden. (in Zeichen: t_d)*
- $\frac{\frac{1}{d}S_X^2}{\frac{1}{m}S_Y^2}$*, eine* **(Fisher-)F-Verteilung** *mit d und m Freiheitsgraden. (in Zeichen $F_{d,m}$)*
- $\mu + \sigma\frac{X_1}{Y_1}$ *heißt* **Cauchy-Verteilung** *mit Lageparameter μ und Streuparameter σ. Diese Verteilung hat weder Erwartungswert noch Varianz.*

Zur jeder dieser Verteilungen gibt es noch nichtzentrale Versionen, wenn die Normalverteilungen in der Angabe einen von Null verschiedenen Mittelwertsparameter besitzen.

Nach dieser Definition gelten offenbar die folgenden Beziehungen:

- Sind $X \sim \chi_n^2$ und $Y \sim \chi_m^2$ unabhängig χ^2-verteilt, so ist $X + Y \sim \chi_{n+m}^2$, da man ja nur mehrere unabhängige Quadratzahlen addiert.
- Ist $X \sim F_{d_1,d_2}$ so ist $\frac{1}{X} \sim F_{d_2,d_1}$, da sich ja nur der Bruch umkehrt.
- Ist $X \sim N(0,1)$, so ist $X^2 \sim \chi_1^2$
- Ist $X \sim t_d$, so ist $X^2 \sim F_{1,d}$

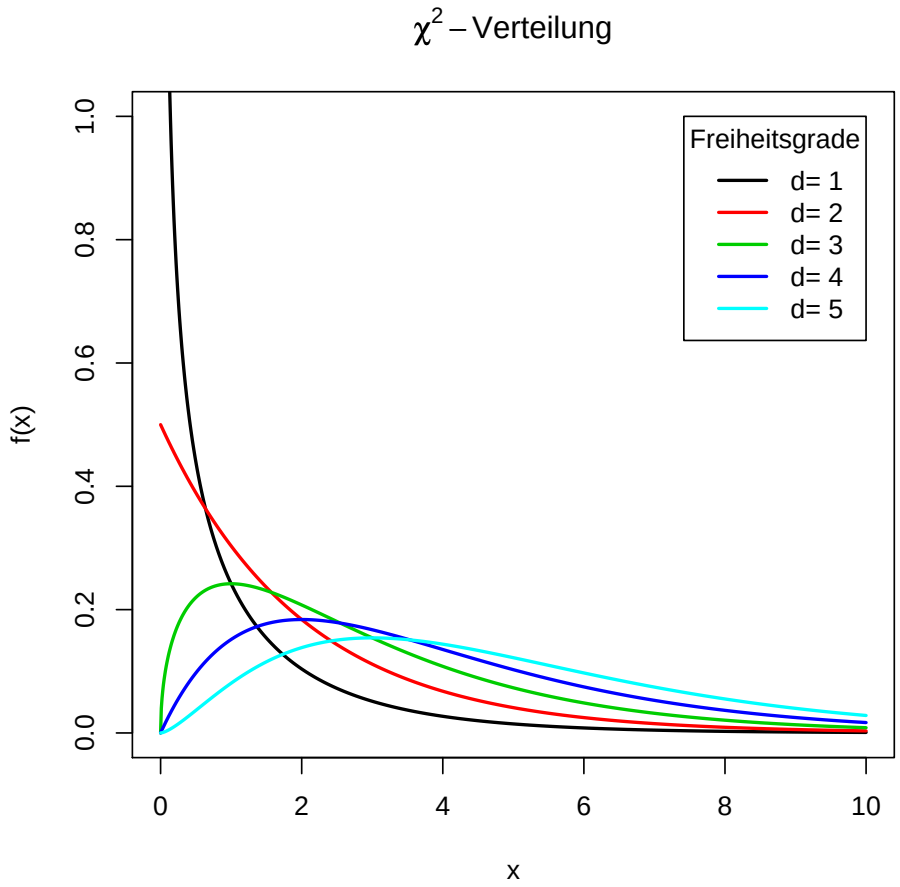

Abbildung 3.5: χ^2-Verteilung

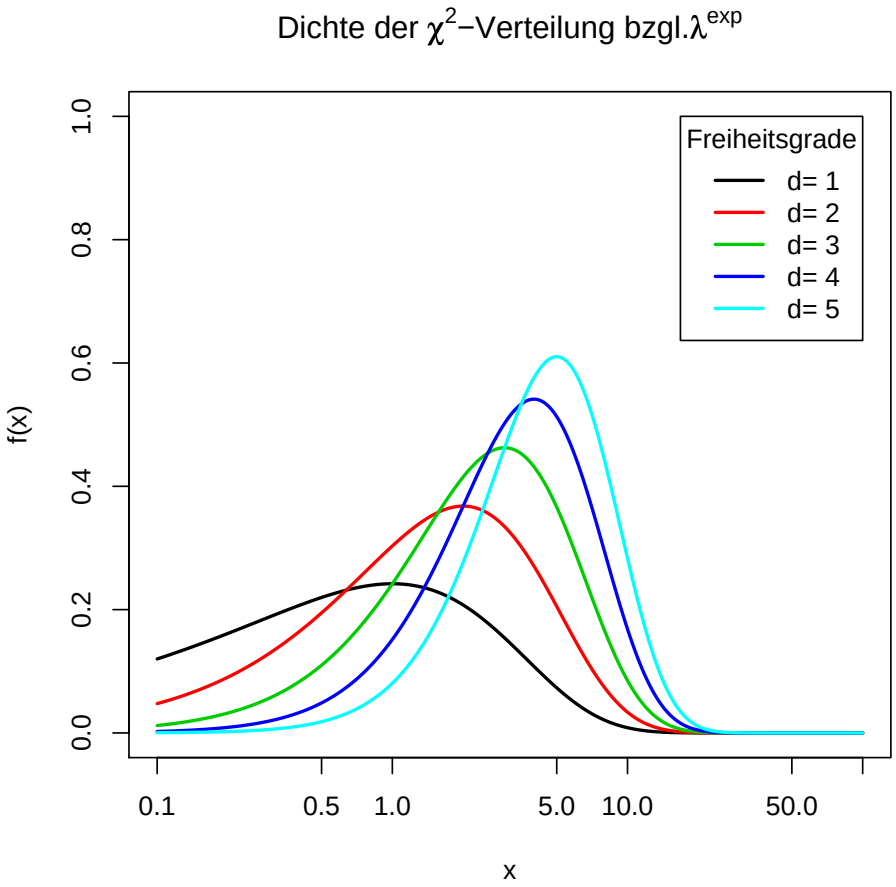

Abbildung 3.6: Dichte des Logarithmus einer χ^2-verteilten Zufallsgröße

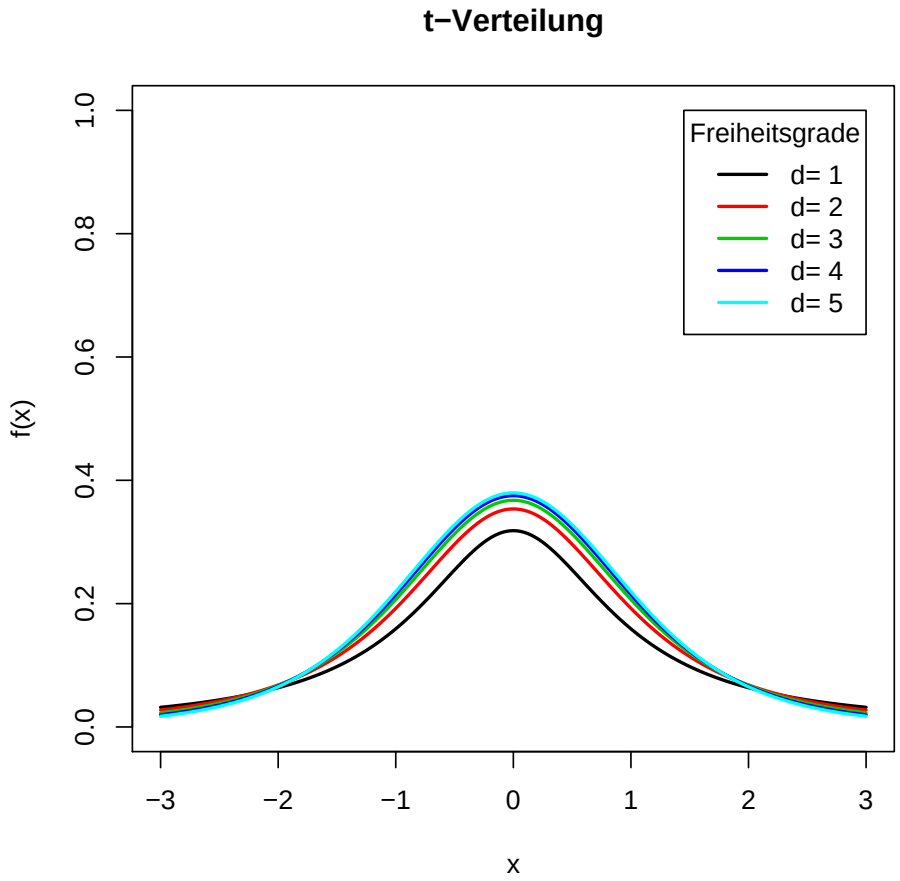

Abbildung 3.7: t-Verteilung

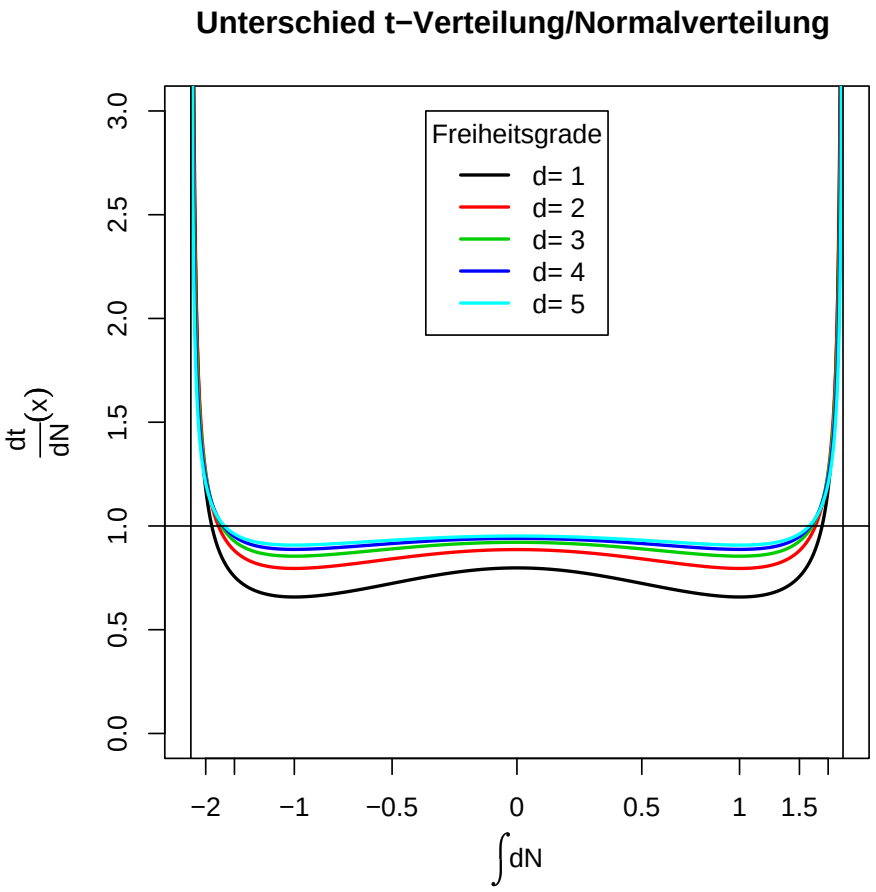

Abbildung 3.8: $N(0,1)$-Dichte der t-Verteilung

- Ist $X \sim N(0,\sigma^2)$ und $Y \sim \sigma^2\chi^2_d$ so ist

$$\frac{X}{\sqrt{\frac{1}{d}Y}} \sim t_d$$

- Ist $X_i \sim N(0,1)$ und $X = (X_i)_{i=1,\ldots,d}$, so ist $\|X\|^2 = \sum_{i=1}^d X_i^2 \sim \chi^2_d$.

Außerdem gilt noch (ohne Beweis):

- Die χ^2_d-Verteilung entspricht der $Gamma(1, \frac{d}{2})$-Verteilung. Man kann χ^2-Verteilungen in dieser Weise auch mit nicht ganzzahligen Freiheitsgraden definieren.

Satz 35 (Satz von Gauss-Markov) *Im Modell $X_i \sim N(\mu,\sigma^2)$ sind $\bar{X}$ und $\hat{\sigma}^2$ stochastisch unabhängig. Dabei ist $\bar{X} \sim N(\mu, \frac{1}{n}\sigma^2)$ verteilt und $\frac{n-1}{\sigma^2}\hat{\sigma}^2$ ist χ^2-verteilt mit $n-1$-Freiheitsgraden.*

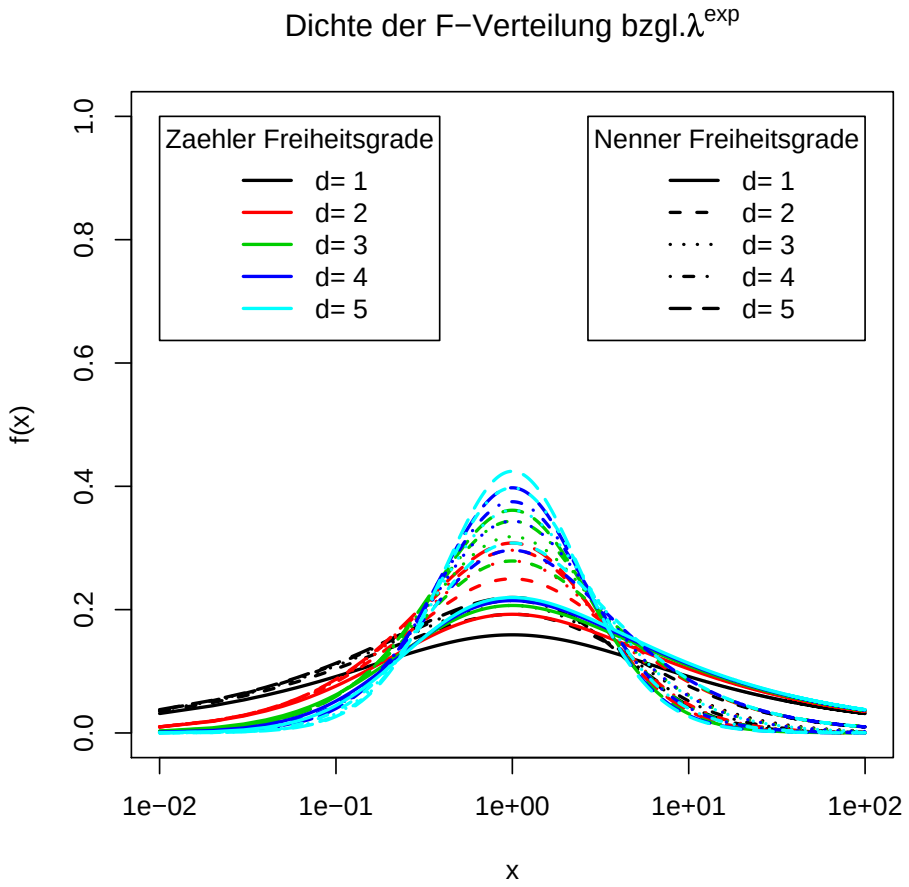

Abbildung 3.9: Dichte der F-Verteilung zu verschiedenen Freiheitsgraden.

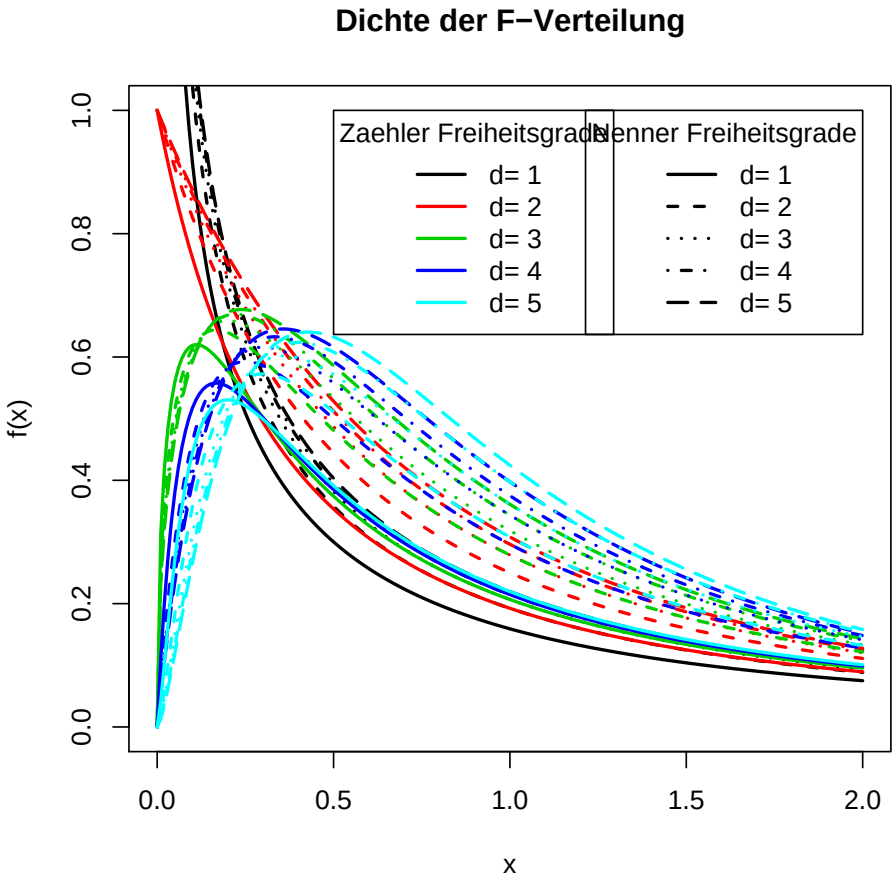

Abbildung 3.10: λ^{exp}-Dichte der F-Verteilung.

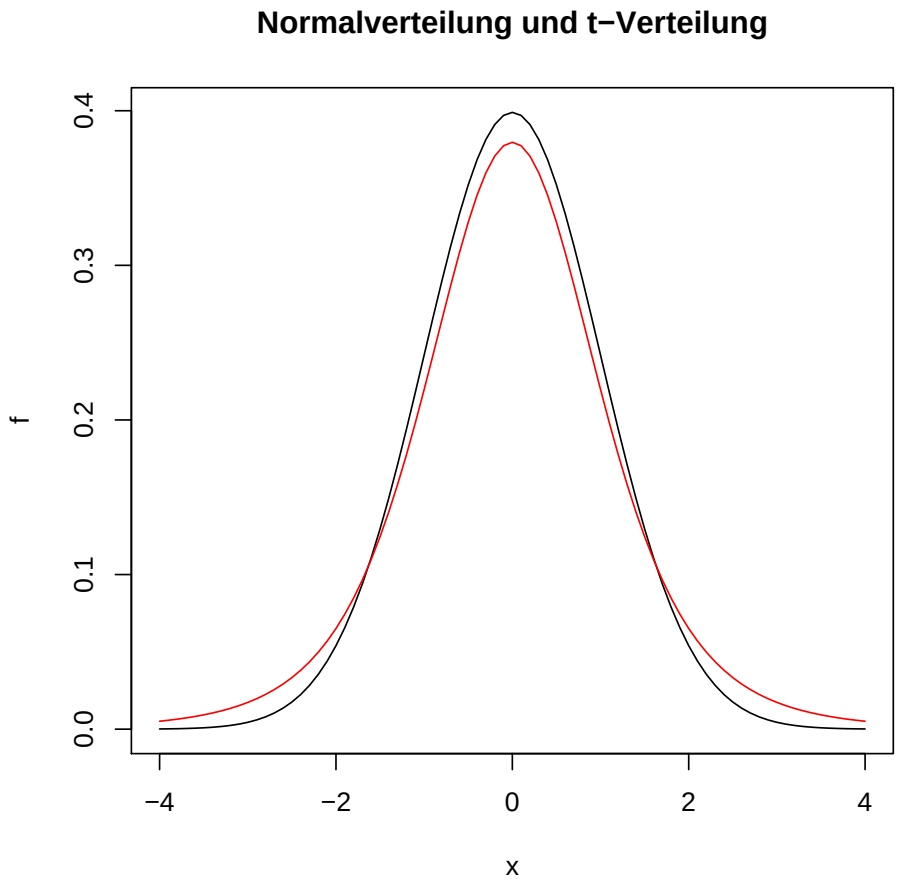

Abbildung 3.11: Vergleich von t-Verteilung und Normalverteilung

Beweis: : Sei $X := (X_i)_{i=1,\ldots,n}$ ein Zufallsvektor. Dieser ist wegen der Unahängigkeit $N(\mathbb{1}_n\mu, \sigma^2\mathsf{Id})$ verteilt. Weiterhin ist $\bar{X} = \frac{1}{n}\mathbb{1}_n^t X = \underbrace{\begin{pmatrix} \frac{1}{n} & \cdots & \frac{1}{n} \end{pmatrix}}_{\mathbf{c}^t} X$ und

$$(X_i - \bar{X}) = X - \mathbb{1}_n \frac{1}{n}\mathbb{1}_n^t X = \underbrace{(\mathsf{Id} - \mathbb{1}_n \frac{1}{n}\mathbb{1}_n^t)}_{:=P} X$$

Mit dieser Setzung für eine Matrix P (P für Projektor) ist $\hat{\sigma}^2 = \frac{1}{n-1}\|X - \mathbb{1}_n \frac{1}{n}\mathbb{1}_n^t X\|^2 = \frac{1}{n}\|PX\|^2$.

P ist ein orthogonaler Projektor im Sinne der lineare Algebra:

$$P^t = \mathsf{Id}^t - \frac{1}{n}\mathbb{1}_n\mathbb{1}_n^t = P$$

$$\begin{aligned} PP &= (\mathsf{Id} - \frac{1}{n}\mathbb{1}_n\mathbb{1}_n^t)(\mathsf{Id} - \frac{1}{n}\mathbb{1}_n\mathbb{1}_n^t) && (3.1)\\ &= \mathsf{Id} - \frac{1}{n}\mathbb{1}_n\mathbb{1}_n^t\mathsf{Id} - \mathsf{Id}\frac{1}{n}\mathbb{1}_n\mathbb{1}_n^t + \frac{1}{n}\mathbb{1}_n \underbrace{\mathbb{1}_n^t\frac{1}{n}\mathbb{1}_n}_{1}\mathbb{1}_n^t && (3.2)\\ &= \mathsf{Id} - \frac{1}{n}\mathbb{1}_n\mathbb{1}_n^t && (3.3)\\ &= P && (3.4) \end{aligned}$$

und $\mathbf{c}$ liegt in seinem Kern:

$$Pc = (\mathsf{Id} - \frac{1}{n}\mathbb{1}_n\mathbb{1}_n^t)\frac{1}{n}\mathbb{1}_n = \frac{1}{n}\mathbb{1}_n - \frac{1}{n}\mathbb{1}_n \underbrace{\mathbb{1}_n^t\frac{1}{n}\mathbb{1}_n}_{1}$$

Ausserdem liegen wegen $Px - x = \mathsf{Id}\mathbf{x} - \frac{1}{n}\mathbb{1}_n\mathbb{1}_n^t\mathbf{x} - \mathbf{x} = \mathbb{1}_n(\frac{1}{n}\mathbb{1}_n^t\mathbf{x})$, nur Vektoren parallel zu $\mathbb{1}_n$ im Kern. Es gilt also $\ker P = span\,\langle c\rangle$.

Wir zeigen nun, dass PX und $\bar{X}$ unabhängig sind. Nach dem Transformationssatz für die Normalverteilung gilt:

$$\begin{pmatrix} PX \\ \bar{X} \end{pmatrix} \sim N\left(\begin{pmatrix} P \\ \mathbf{c}^t \end{pmatrix} X, \begin{pmatrix} P \\ \mathbf{c} \end{pmatrix} \mathsf{Id} \begin{pmatrix} P^t & \mathbf{c} \end{pmatrix}\right)$$

und somit wegen

$$\begin{pmatrix} P \\ \mathbf{c}^t \end{pmatrix} \sigma^2\mathsf{Id} \begin{pmatrix} P^t & \mathbf{c} \end{pmatrix} = \sigma^2 \begin{pmatrix} PP^t & Pc \\ \mathbf{c}^t & \mathbf{c}^t\mathbf{c} \end{pmatrix} = \sigma^2 \begin{pmatrix} P & 0 \\ 0 & \frac{1}{n} \end{pmatrix}$$

und

$$\begin{pmatrix} P \\ \mathbf{c}^t \end{pmatrix} \mathbb{1}_n\mu = \begin{pmatrix} Pn\mathbf{c}\mu \\ \frac{1}{n}\mathbb{1}_n^t\mathbb{1}_n\mu \end{pmatrix} = \begin{pmatrix} 0 \\ \mu \end{pmatrix}$$

gilt somit

$$\begin{pmatrix} PX \\ \bar{X} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ \mu \end{pmatrix} X, \sigma^2 \begin{pmatrix} P & 0 \\ 0 & \frac{1}{n} \end{pmatrix}\right)$$

Beide PX und $\bar{X}$ sind somit stochastisch unabhängig und damit auch $\hat{\text{var}}(X)$ und $\bar{X}$.

Für die χ^2_{n-1}-Verteilung von $\|PX\|^2 = \sum_{i=1}^n (X_i - \bar{X})^2$ muss man ein Stück tiefer in die lineare Algebra einsteigen. Für P gilt

- Da P ein orthogonaler Projektor ist, hat er nur die Eigenwerte 0 und 1.

- P als symmetrische Matrix ist durch eine orthogonale Matrix Q diagonalisierbar:

$$P = Q^t \underbrace{\begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix}}_{=:D} Q$$

- Da der Kern eindimensional ist, gilt für die Eigenwerte $\lambda_1, \ldots, \lambda_{n-1} = 1$ und $\lambda_n = 0$

Also gilt für $Y := \frac{1}{\sigma} QPX$

$$\|PX\|^2 = \|QPX\|^2 = \sigma^2 \|Y\|^2$$

Berechnen wir aber die Verteilung von Y so erhalten wir

$$Y \sim N(Q0, \frac{1}{\sigma^2}\sigma^2 QQ^t DQQ^t) = N(0, D)$$

und somit $Y_i \sim N(0,1), i.i.d.$ für $i = 1, \ldots, n-1$ und $Y_n \equiv 0$ P-fast sicher. und somit $\frac{1}{\sigma^2}\|PX\|^2 \sim \chi^2_{n-1}$ □

Mit den Bezeichnungen der Definition 34 gilt also: $\bar{X} - \mu$ ist verteilt wie $\sqrt{\frac{\sigma^2}{n}} Y_1$ und $\hat{\sigma}^2$ ist verteilt wie $\frac{\sigma^2}{n-1} S_X^2$ und beide sind unabhängig, also entsprechen sich auch die gemeinsamen Verteilungen.

Daher gilt:

$$S(X) = \frac{\bar{X} - \mu}{\sqrt{\frac{1}{n}\hat{\sigma^2}}} \sim \frac{\sqrt{\frac{\sigma^2}{n}} Y_1}{\sqrt{\frac{1}{n}\frac{\sigma^2}{n-1} S_X^2}} = \frac{Y_1}{\sqrt{\frac{1}{n-1} S_X^2}} \sim t_{n-1}$$

und Teststatistik $S(X)$ ist somit t-verteilt.

Der Ein-Stichproben-t-Test kann also ansonsten analog zum Ein-Stichproben Gauss-Test verwendet werden.

Beispiel 36 *Ein Ein-Stichproben-t-Test wird z.B. eingesetzt, wenn nachgewiesen werden soll, dass ein mehrfach gemessener Wert (mit normalverteiltem Messfehler) einen Grenzwert nicht übersteigt.*

3.3.3 Die Zwei-Stichproben-t-Tests

Beim Zwei-Stichproben-t-Test wird der Erwartungswert nicht mit einem festen Wert sondern mit dem Erwartungswert einer anderen Grundgesamtheit verglichen:

$$H_0 : \mu_X = \mu_Y \text{ vs. } H_0 : \mu_X \neq \mu_Y$$

Dazu benötigt man Stichproben aus beiden Grundgesamtheiten:

$$X_i \sim N(\mu_X, \sigma^2), i = 1, \ldots, n, \ i.i.d.$$

und unabhängig davon

$$Y_i \sim N(\mu_Y, \sigma^2), i = 1, \ldots, m, \ i.i.d.$$

Anstelle des Mittelwertes minus dem Vergleichswert tritt nun die Differenz der Mittelwerte:

$$Z = \bar{X} - \bar{Y}$$

Als Linearkombination normalverteilter Größen ist Z wieder normalverteilte und zwar unter der Hypothese mit dem Erwartungswert $E[Z] = E[\bar{X}] - E[\bar{Y}] = \mu_X - \mu_Y = 0$ und unter der Voraussetzung der gleichen Varianz in beiden Grundgesamtheiten mit der Varianz

$$\mathrm{var}(Z) = \frac{1}{n}\mathrm{var}(X) + \frac{1}{m}\mathrm{var}(Y) = \left(\frac{1}{n} + \frac{1}{m}\right)\sigma^2$$

Auf der anderen Seite ist wieder nach Gauss-Markov

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \bar{X}) \sim \chi^2_{n-1}$$

und

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i - \bar{Y}) \sim \chi^2_{m-1}$$

und somit

$$N := \sum_{i=1}^{n}(X_i - \bar{X}) + \sum_{i=1}^{n}(Y_i - \bar{Y}) \sim \sigma^2\chi^2_{(n-1)+(m-1)}$$

Die entsprechend normierte Teststatistik:

$$t := \frac{Z}{\sqrt{\left(\frac{1}{n} + \frac{1}{m}\right)\frac{n+m}{N}}} = \frac{\bar{X} - \bar{Y}}{\sqrt{\left(\frac{1}{n} + \frac{1}{m}\right)\frac{n+m}{\sum}\sum_{i=1}^{n}(X_i - \bar{X}) + \sum_{i=1}^{n}(Y_i - \bar{Y})}}$$

folgt also wieder einer t-Verteilung mit $n+m-2$ Freiheitsgraden, da beide Nennerteile von beiden Zählerteilen nach Gauss-Markov unabhängig sind.

- **Zwei-Stichproben-t-Test**

  Situation: Vergleich von Mittelwerten zweier Stichproben bei Normalverteilung und gleicher Varianz

  $H_0 : \mu_X = \mu_Y$

  $H_1 : \mu_X \neq \mu_Y$ oder $\mu_X > \mu_Y$ oder $\mu_X < \mu_Y$

  Voraussetzungen: $X_i \sim N(\mu_X, \sigma^2)$ und $Y_i \sim N(\mu_Y, \sigma^2)$, i.i.d.

  Bemerkung: Die Normalverteilungsvoraussetzung ist relativ unkritisch, solange keine Ausreißer vorliegen und die Verteilung ungefähr normal ist. Die Normalverteilungsvoraussetzung kann mit dem Shapiro-Wilk Test und die Varianzgleichheit mit dem F-Test überprüft werden.

```
> x <- rnorm(10, mean = 2, sd = 4)
> y <- rnorm(10, mean = 3, sd = 4)
> shapiro.test(x)
> shapiro.test(y)
> var.test(x, y)
> t.test(x, y, alternative = "two.sided")
```

Beispiel 37 *Ein Zwei-Stichproben-t-Test wird z.B. eingesetzt, wenn nachgewiesen werden soll, dass ein Gen in Krebszellen anders exprimiert ist, als in gesunden Zellen. Dazu werden Messwerte der Expression aus einer Stichprobe mit Krebszellen und aus einer Stichprobe von gesunden Zellen benötigt. Vor dem Einsatz ist die Voraussetzung der Normalverteilung zu prüfen.*

Beispiel 38 *Ein Zwei-Stichproben-t-Test wird z.B. eingesetzt, wenn nachgewiesen werden soll, dass Männer und Frauen bei einer bestimmten Qualifikation (z.B. Berufseinsteiger als Diplom-Geologe) unterschiedlich viel verdienen.*

3.3.4 Welch-t-Test

Die Voraussetzung der gleichen Varianz ist natürlich etwas leichtgläubig. Wenn man sich schon nicht sicher ist, dass die Populationen den gleichen Erwartungswert haben, wie kann man dann gesichert annehmen sie haben die gleiche Varianz. Ein realistischeres Modell wäre also, dass die Varianzen verschieden sind. Dazu benötigt man Stichproben aus beiden Grundgesamtheiten:

$$X_i \sim N(\mu_X, \sigma_X^2), i = 1, \ldots, n, \ i.i.d.$$

und unabhängig davon

$$Y_i \sim N(\mu_Y, \sigma_Y^2), i = 1, \ldots, m, \ i.i.d.$$

Für dieses Problem hat B.L. Welch 1947 einen modifizierten Test vorgeschlagen.
Bei stark ungleichen Varianzen $\sigma_X^2 \not\sigma_Y^2$ ergeben sich zwei Probleme.

- Der Erwartungswert der Varianzschätzung im Nenner wird:

  $$(\frac{1}{n} + \frac{1}{m})\frac{n\sigma_X^2 + m\sigma_Y^2}{n+m}$$

  während die Zählervarianz zu

  $$\frac{1}{n}\sigma_X^2 + \frac{1}{m}\sigma_Y^2$$

  wird. Ist dabei z.B. $m \gg n$ so bedeutet das, dass das σ_Y^2 im Zähler relativ heruntergewichtet wird und im Nenner relativ heruntergewichtet wird. Um das zu vermeiden verwendet Welch eine modifizierte Teststatistik

  $$t_W := \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{1}{n}\hat{\sigma}_X^2 + \frac{1}{m}\hat{\sigma}_X^2}}$$

  in der die im Nenner geschätzte Varianz der Zählervarianz entspricht.

- Zum Zweiten wird dadurch natürlich in Abhängigkeit vom tatsächlichen Verhältnis der Varianzen die Varianz verschieden genau geschätzt. Dominiert z.B. σ_Y^2, so ist die Schätzung ungefähr so genau, wie wenn man nur σ_Y^2 geschätzt hätte. Sind jedoch beide Varianzen tatsächlich gleich und $n = m$ so unterscheidet sich die Schätzung nicht vom Zwei-Stichproben-t-Test. Um dieses Problem zu umgehen, hat Welch vorgeschlagen die Freiheitsgrade mit einer vom Verhältnis der Varianzen abhängigen Zahl zu berechnen (Formel von Welch-Satterthwaite):

  $$d = \frac{\frac{\hat{\sigma}_X^2}{n} + \frac{\hat{\sigma}_Y^2}{M}}{\frac{\hat{\sigma}_X^4}{n^2(n-1)} + \frac{\hat{\sigma}_X^4}{m^2(m-1)}}$$

Das hat verschiedene Konsequenzen:

- Die Teststatistik erfüllt die Verteilung also nur approximativ.
- Je nach dem wie die Varianzen geschätzt werden, wird mit einer anderen Verteilung verglichen. Solche Tests die Verteilungen benutzen, die bedingt an einer Nebenbeobachtung sind (hier bedingt an der geschätzten Varianz) nennt man **bedingte Tests**.

Welchs t–Test

Situation: Vergleich von Mittelwerten zweier Stichproben bei Normalverteilung und verschiedener Varianz

$H_0 : \mu_X = \mu_Y$

$H_1 : \mu_X \neq \mu_Y$ oder $\mu_X > \mu_Y$ oder $\mu_X < \mu_Y$

Voraussetzungen: $X_i \sim N(\mu_X, \sigma_X^2)$ und $Y_i \sim N(\mu_Y, \sigma_Y^2)$, i.i.d.

Bemerkung: Die Normalverteilungsvoraussetzung ist relativ unkritisch, solange keine Ausreißer vorliegen und Verteilung ungefähr normal ist.

3.3.5 Überprüfen der Varianzgleichheit

Zur Überprüfung der Varianzgleichheit bietet es sich an als Teststatistik die beiden Varianzschätzer durcheinander zu teilen:

$$F := \frac{\sigma_X^2}{\sigma_Y^2}$$

unter der Nullhypothese des Testsproblems:

$$H_0 : \sigma_X^2 = \sigma_Y^2 \text{ vs. } H_1 : \sigma_X^2 \neq \sigma_Y^2$$

ist dieses Statistik nach Kürzen von σ_X^2 offenbar F-Verteilt mit $n-1$ und $m-1$ Freiheitsgraden, da Zähler und Nenner ja wegen der stochastischen Unabhängigkeit der Stichproben jeweils unabhängig sind und nach Gauss-Markov mit einer mit σ_X^2 skalierten χ^2-Verteilung verteilt sind.

- **F-Test**

  Situation: Test auf Gleichheit der Varianz bei Normalverteilung

  $H_0 : \sigma_X^2 = \sigma_Y^2$

  $H_1 : \sigma_X^2 \neq \sigma_Y^2$ oder $\sigma_X^2 > \sigma_Y^2$ oder $\sigma_X^2 < \sigma_Y^2$

  Voraussetzungen: $X_i \sim N(\mu, \sigma_X^2)$, $Y_i \sim N(\mu, \sigma_Y^2)$

  Bemerkung:

```
> var.test(x, y)
```

Entsprechend kann man natürlich auch mit einer festen Referenzvarianz σ_0^2

$$H_0 : \sigma_X^2 = \sigma_0 \text{ vs. } H_1 : \sigma_X^2 \neq \sigma_0^2$$

vergleichen, indem man direkt die nach dem Satz von Gauss-Markov χ_{n-1}^2 verteilte Teststatistik

$$\chi^2 = \frac{(n-1)\hat{\sigma}_X^2}{\sigma_0^2}$$

verwendet.

3.3.6 Gepaarter t-Test

Ein häufiges Testproblem ist der Vergleich von zwei Messwerten an der selben Person, z.B. wenn man Zustand X_i vor und Y_i nach einer Behandlung vergleichen möchte. Für den gepaarten t-Test, der dieses Problem bearbeitet, muss man voraussetzen, dass

$$\Delta_i = Y_i - X_i \sim N(\mu, \sigma^2)$$

und die Δ_i unabhängig identisch mit gleicher Varianz verteilt sind. Das Testproblem lautet dann

$$H_0 : E[\Delta_i] = 0 \text{ vs. } E[\Delta_i] \neq 0$$

Diese Testproblem kann man offenbar mittels eines Ein-Stichproben-t-Test für Δ_i lösen. Weil der Test aber viel häufiger in diesem Gewandt paarweiser Beobachtungen auftritt, bekommt dieser Test einen eigenen Namen und heißt gepaarter t-Test, weil er sich auf Paare stochastisch abhängiger Beobachtungen bezieht.

> Vergleicht man stochastisch abhängige Beobachtungspaare, so benötigt man einen gepaarten Test. Vergleicht man hingegen zwei stochastisch unabhängige Stichproben, so braucht man einen Zwei-Stichproben-Test.

3.3.7 Normalverteilungsmodelle mit mehreren Stichproben

Der zur Familie der t-Tests gehörende Mehrstichprobentest zum Vergleich von Mittelwerten heißt Varianzanalyse. Die Varianzanalyse wird im Rahmen der linearen Modelle besprochen:

- **Einfache Varianzanalyse**

  Situation: Test auf Gleichheit der Erwartungswerte mehrerer normalverteilter Stichproben.

  $H_0 : \forall g, g' : \mu_g = \mu_{g'}$

  $H_1 : \exists g, g' : \mu_{g_i} \neq \mu_j$

  Voraussetzungen: $X_i \sim N(\mu_{g_i})$ wobei g_i die Gruppenzugehörigkeit des Individuums i beschreibt.

  Bemerkung: Die Varianzanalyse setzt die Gleichheit der Varianz und Normalverteilung voraus.

```
> data(iris)
> anova(lm(Sepal.Length ~ Species, data = iris))
```

  Der p-Wert kann unter `P(>F)` abgelesen werden.

Der entsprechende Test zum Vergleich der Varianzen mehrerer Stichproben heißt Bartlett-Test:

- **Bartlett–Test**

  Situation: Testet auf gleiche Varianz in mehrere Stichproben

  H_0 : Die Varianzen der Stichproben sind gleich.

  H_1 : Die Varianzen der Stichproben sind nicht alle gleich.

  Voraussetzungen: $X_i|G_i \sim N(\mu_{G_i}, \sigma_{G_i})$ stochastisch unabhängig.

  Bemerkung: Dieser Test wird oft eingesetzt, um eine Voraussetzung der Varianzanalyse zu überprüfen.

```
> bartlett.test(Sepal.Length ~ Species, data = iris)
```

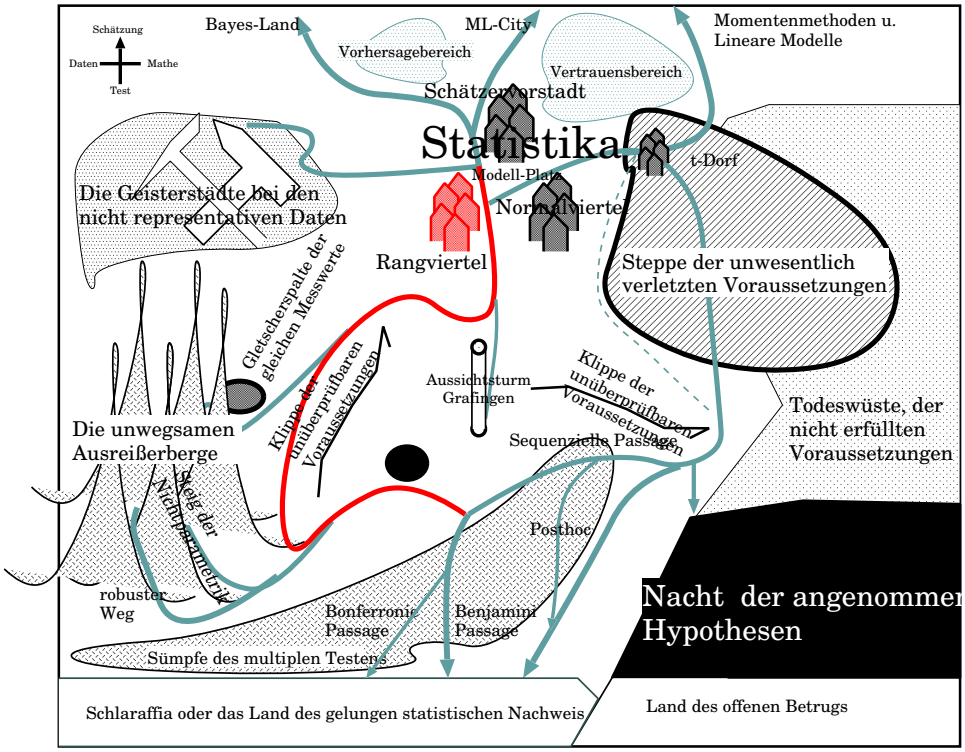

Abbildung 3.12: Übersichtskarte Rangbasierte Tests
Die Einordnung von Kapitel 3.4 in die Übersichtskarte.

3.4 Die Rangbasierten Tests

Die im Abschnitt 3.3 besprochenen Test setzen sämtlich die Normalverteilung der beteiligten Größen voraus. Das hat Gründe:

- Nur wenn die Verteilung der Teststatitik unter der Nullhypothese eindeutig beschrieben werden kann, oder wenigstens eine Obergrenze für die Quantile angegeben werden kann, ist es möglich den kritischen Wert für den Test zu berechnen. Man benötigt also eigentlich ein konkretes Verteilungsmodell.

- Setzt man Normalverteilung voraus, so lassen sich diese Verteilungen besonders leicht berechnen, insbesondere auch, da dann der Satz von Gauss-Markov gilt. Es ist also leicht mit Normalverteilung zu arbeiten.

Dennoch ist die Voraussetzung der Normalverteilung in vielen Fällen inkorrekt. Wir benötigen daher eine Klasse von Tests, die von dieser Voraussetzung unabhängig funktioniert. Dabei gibt es eine einfache Idee, die wenigstens den ersten der beiden oben genannten Punkte erfüllt: Verwendet man anstelle der Daten ihre Ränge, so sind die sich ergebenden Verteilungen oft unabhängig von der konkreten Verteilungsfunktion. Das gilt allerdings nur, wenn die Wahrscheinlichkeit für doppelt auftretende Werte verschwinden gering ist (eine P-Nullmenge).

3.4.1 Wilcoxen-Rang-Summen-Test

Denken wir zurück an die Situation des Zwei-Stichproben-t-Test:

$$H_0 : \mu_X = \mu_Y \text{ vs. } H_0 : \mu_X \neq \mu_Y$$

Jetzt wollen wir aber keine Normalverteilung voraussetzen, sondern sehen in μ einen beliebigen Lageparameter. Wir setzen also voraus, dass X und Y die gleiche grundsätzliche Verteilung haben, diese nur einmal in μ_X und einmal in μ_Y verschoben ist:

$$F_X(x) = F(x - \mu_X),\ F_Y(x) = F(x - \mu_Y)$$

für irgend eine (unbekannte aber stetige) Verteilungsfunktion F. Unter den folgenden Voraussetzungen:

- Die Hypothese gilt
- Die X_i und Y_i sind alle stochastisch unabhängig voneinander
- F ist wirklich stetig

hängt die Verteilung der gemeinsamen Ränge (rang X_i,rang Y_i) der X_i und Y_i überhaupt nicht von F oder μ_X oder μ_Y ab, sondern ist einfach die Gleichverteilung auf der Menge der möglichen Anordnungen. Die Verteilung einer beliebigen Teststatistik, die nur auf solchen Rängen basiert, kann also (zumindest im Prinzip, auch wenn wir das nicht wirklich selber machen wollen) berechnet werden.

Der Wilcoxen-Rang-Summen-Test verwendet dabei die einfache Teststatistik:

$$W = \sum_{i=1}^{n} \text{rang}\, X_i$$

Diese nimmt offenbar tendenziell große Werte an, wenn die X_i über den Y_i liegen und kleine Werte, wenn die X_i unter den Y_i liegen.

- **Wilcoxen–Rang–Summen–Test**

  Situation: Vergleich der Lage zweier Stichproben mit stetiger Verteilung

  $H_0 : \forall x : F_X(x) = F_Y(x)$

  $H_1 : \forall x : F_X(x) = F_Y(x - c)$ mit $c \neq 0$ oder $c > 0$ oder $c < 0$

  Voraussetzungen: X_i und Y_i sind alle stochastisch unabhängig und die F_X und F_Y sind stetig.

  Bemerkung: Es handelt sich um ein rangbasiertes Verfahren. Der Test wird allgemein verwendet um die Lagegleichheit bei nicht normalverteilten Stichproben zu testen, da die Voraussetzung der Verteilungsgleichheit für die Korrektheit des Tests unkritisch ist. Der Test wird ungenau, wenn gleiche Werte (Bindungen) vorkommen.

```
> x <- rexp(10, 5)
> y <- rnorm(12, 3)
> wilcox.test(x, y)
```

3.4.2 Wilcoxen-Vorzeichen-Rang-Test

Eine Entsprechung für den gepaarten t-Test findet sich im Wilcoxen-Vorzeichen-Rang-Test. Die Hypothese, dass zwischen X_i und Y_i keine Änderung stattgefunden hat wird ersetzt durch die Idee, dass X_i und Y_i austauschbar sind und damit $\Delta_i = Y_i - X_i$ symmetrisch um 0 verteilt sein sollte. Ist Y_i im Mittel um einen festen Betrag verschoben, so sollte Δ_i um den entsprechenden Verschiebungswert symmetrisch verteilt sein.

Das Testproblem lautet also wieder

$$H_0 : \mu = 0 \text{ vs. } \mu \neq 0$$

aber jetzt unter der allgemeineren Voraussetzung $F_\Delta(x-\mu) = 1 - F_\Delta(-(x-\mu))$.

Wilcoxen verwendet hierfür eine Teststatistik, die auf den Rängen von rang $|\Delta_i|$ und den Vorzeichen $\operatorname{sgn} \Delta_i = \pm 1$ der Δ_i beruht

$$W := \sum_{i=1}^{n} \operatorname{sgn} \Delta_i \operatorname{rang} |\Delta_i|$$

- **Wilcoxen-Vorzeichen-Rang-Test**

  Situation: Testet auf eine mittlere Änderung von 0 zwischen beiden Beobachtungen am gleichen Individuum.

  H_0 : Die Verteilung von $X_i - Y_i$ ist symmetrisch um 0.

  H_1 : Die Verteilung von $X_i - Y_i$ ist symmetrisch um ein $c \neq 0$ oder $c < 0$ oder $c > 0$

  Voraussetzungen: Die Verteilung ist für alle Paare gleich.

  Bemerkung: Dieses rangbasierte Verfahren hat Probleme mit Bindungen in den Differenzen.

```
> x <- rcauchy(10, 10, 5)
> y <- x + rcauchy(10, 0.1, 0.05)
> wilcox.test(x, y, paired = TRUE)
```

3.4.3 Weitere rangbasierte Tests

Weitere wichtige rangbasierte Verfahren sind:

- Der Kruskal-Wallis-Test und der Witney-U-Test, welche Mehrstichprobenentsprechung des Wilcoxen-Rang-Summen-Tests bilden.
- Der Fligner-Test, welcher die Hypothese gleicher Streuung für zwei und mehr Stichproben testen kann.

3.5 Anpassungstest

Anpassungstests sind Ein-Stichproben-Tests, welche die Passgenauigkeit der Verteilung einer Stichprobe zu einer vorgegebene Verteilung oder Verteilungsfamilie testen. Die wichtigsten Anpassungstests sind:

- Der **Shapiro-Wilk-Test**
  Der Shapiro-Wilk-Test eignet sich zum Vergleich mit einer Normalverteilung:

  $$H_0 : \text{X ist normalverteilt vs.} H_1 : \text{X ist nicht normalverteilt}$$

  Grob gesprochen testet der Shapiro-Wilk-Test, ob der QQ-Plot gerade genug aussieht.

- Der **Kolmogorov-Smirnov-Test**
  Der Kolmogorov-Smirnov-Test eignet sich zum Vergleich mit einer beliebigen, aber festen Verteilung.

  $$H_0 : \text{X ist } P_0\text{-verteilt vs.} H_1 : \text{X ist nicht } P_0\text{-verteilt}$$

  Die Teststatistik ist der maximale Abstand zwischen empirischer und theoretischer Verteilungsfunktion.

- Der χ^2-Anpassungstest
  Der χ^2-Anpassungstest ist eine stumpfe Allzweckmethode, die nur im Rahmen der asymptotischen Statistik richtig verstanden werden kann.

3.6 Multiples Testen

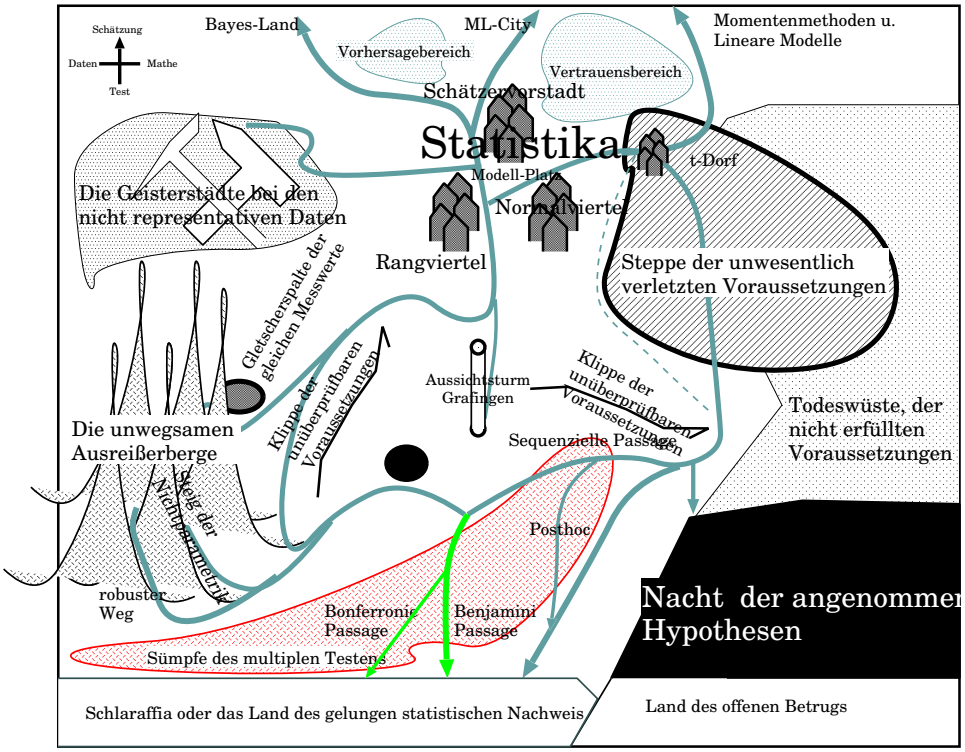

Abbildung 3.13: Übersichtskarte Multiples Testen
Die Einordnung von Kapitel 3.6 in die Übersichtskarte.

Auswegen aus dem multiplen Testen:

- Sequenzielles Testen I:
  Ein weiterer Test wird nur durchgeführt, wenn der vorherige schon signifikant war.

- Sequenzielles Testen II:
  Vor der eigentliche Analyse wird ein Test gemacht, der keine wissenschaftlich inhaltliche Aussage prüft, sondern lediglich entscheidet welche Methode angewendet wird. Dabei darf der Test allerdings nicht die jeweils für die Situation leistungsfähigere Methode auswählen, sondern darf lediglich über die Zulässigkeit entscheiden.

- Bonferronie Korrektur:
  Statt p wird Np mit α verglichen, wobei N die Anzahl der durchgeführten Tests ist. Mit dieser Korrektur sind alle signifikaten Ergebnisse statistisch nachgewissen.

- Benjamini-Hochberg-Korrektur (False-Discovery-Rate Ansatz):
  Sind schon einige Tests nach Bonferronie signifikant, so kann man den jeweils nächst größeren Bonferronie korrigierten p-Wert durch die Anzahl der kleineren p-Werte + 1 teilen. Auf diese Weise garantiert man, dass mit $1 - \alpha$ Wahrscheinlichkeit im Mittel höchstens der Anteil α der Hypothesen falsch signifikant ist (d.h. zuzätzlich zum α-Fehler).

3.6.1 Diskussion

3.6.1.1 Verifikation vs. Falsifikation

In den empirischen Wissenschaften ist es grundsätzlich unmöglich Modelle zu verifizieren, da die Welt ja immer wesentlich komplizierter sein kann, als wir sie uns vorstellen.

Diskussion 39 *Verifikation in der Wissenschaft*

- *Die newtonsche Mechanik galt als sicher, bis die Relativitätstheorie aufgestellt wurde.*
- *Z.B. könnte es ein großes unsichtbareres Wissenschaftsmoster geben, dass alle gemachten Experimente bestimmt und zwar immer so dass die Ergebnisse auf irgend etwas hindeuten wozu das Wissenschaftsmoster gerade Lust hat. Die These, dass so ein Wissenschaftsmoster existiert, ist natürlich keine wissenschaftliche, da diese These prinzipiell nicht falsifizierbar ist. Verifiziert ist diese absurde Hypothese natürlich auch nicht. Wäre es aber so, so könnte unser Monster Lust haben, uns sogar die wildesten Hypothesen immmer wieder experimentell zu bestätigen.*
- *Könnte man etwas verifizieren, wenn man voraussetzt, dass kein unsichtbares Wissenschaftsmonster existiert.*
- *Setzt man „***tres non datur***" (ein drittes ist nicht gegeben) voraus, so könnte man doch durch Falsifikation das einer These ihr logisches Gegenteil verifizieren. Gibt es einen qualitativen Unterschied zwischen einer wissenschaftlichen These und ihrem logischen Gegenteil?*

Die Möglichkeit zur Falsifikation ist also eine zentrale Methode der empirischen Wissenschaften.

3.6.1.2 Test vs. Entscheidung vs. Wissen

Viele Menschen haben Schwierigkeiten die Grundvorstellungen der Testtheorie zu akzeptieren, weil Sie intuitiv versuchen das Testproblem als ein Entscheidungsproblem oder ein deterministisches Feststellen zu verstehen, also als ein Versuch die Frage zu beantworten, ob man so handeln sollte, als ob die Hypothese oder als ob die Alternative richtig wäre. Dieses Problem behandelt jedoch die Entscheidungstheorie.

Die Theorie der Tests beschäftigt sich aber mit dem empirischen Nachweis, also einem wissenschaftlichen Konzept. Ein dazu entgegengesetztes Konzept ist das Konzept der statistischen Entscheidung, die zwischen zwei oder mehr Handlungsalternativen auswählt und daher in erster Linie an „guten" Entscheidungen und nicht an „wahren" Aussagen interessiert ist. Die **Entscheidungstheorie** verfolgt ganz andere Ziele als die Testtheorie und ähnelt in ihrem Methoden eher der Schätztheorie. Sie wird ähnlich den Ansätzen der Schätztheorie auf Verlustfunktionen aufgebaut und versuchen den erwarteten Verlust zu minimieren.

Die Entscheidungstheorie antwortet also auf die Frage:

> Soll ich annehmen H_0 stimmt oder soll ich annehmen H_1 stimmt.

entweder mit

> Es lohnt sich H_0 zu vermuten.

oder mit

Es lohnt sich H_1 zu vermuten.

oder vielleicht auch mit:

Es lohnt sich mehr Daten zu sammeln.

Im Gegensatz dazu antwortet die Testtheorie auf die Frage:

Kann H_0 falsifiziert werden (wenn auch H_1 stimmen könnte)?

Mit

H_0 ist (im Rahmen der akzeptierten Unsicherheit) widerlegt.

oder mit

H_0 ist mit dem vorliegenden Datenmaterial nicht zu wiederlegen.

Wir würden uns natürlich eine Supermethodik wünschen, welche die Wahrheit der Hypothese einfach feststellt und auf die Frage:

Gilt H_0 oder H_1?

entweder antwortet:

Es gilt H_0.

oder antwortet

Es gilt H_1.

Aber diese Supermethode gibt es leider nicht und die kann es auch nicht geben.

Die Entscheidungtheorie wird hier nicht näher behandelt.