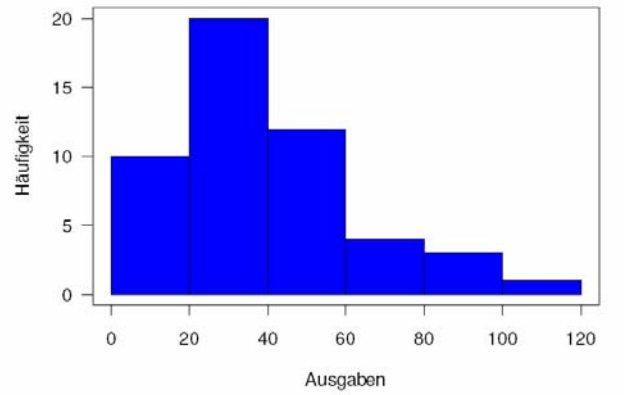
Teil 1: Allgemeine statistische Datenanalyse

1. Stetige Zufallsvariable, Verteilungsfunktion, Quantile, VaR (Value at Risk)

- **Ausgangsbeispiel**
  - Abb. 1 zeigt die Beobachtungsergebnisse von 50 Kunden eines Lebensmittelladens, dargestellt in einem **Rechteckdiagramm.** Die Grösse die interessierte, war, wie viel jeder Kunde ausgab.
  - **Interpretation**
    - Die Mehrheit der Kunden, nämlich 20 Kunden, gaben zwischen 20 und 40 Franken aus.
    - linkssteile/rechtsschiefe Verteilung
    - Die Ausgaben liegen zwischen 0 und 120 Franken.

Abb. 1

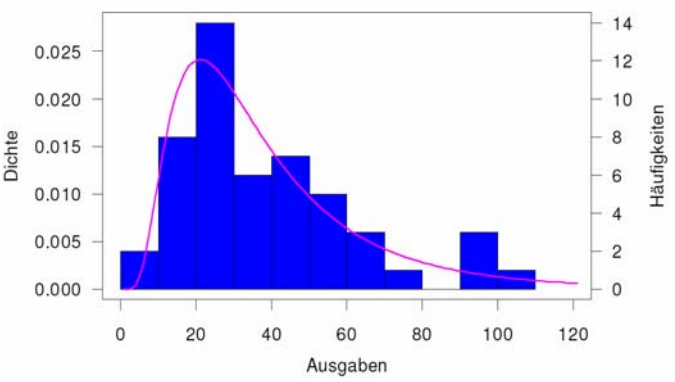


- **1.1 Stetige Zufallsvariable und Dichtefunktion f**
  - Die Ergebnisse einer Beobachtung, die uns interessiert, sind meistens Zahlenwerte. Im Wahrscheinlichkeitsmodell steht das Elementarereignis für die Beobachtungseinheit, während die Beobachtung eines interessierenden Zahlenwertes **Zufallsvariable (ZV)** genannt wird. Der beobachtete Zahlenwert wird **Realisierung** der Zufallsvariablen genannt.
    - **Als Synonyme verwendete Begriffe in der Wahrscheinlichkeitsrechnung**

Zufallsvariable X	Merkmal X
Wahrscheinlichkeit, Realisierung	Relative Häufigkeit
Wahrscheinlichkeitsfunktion	Einfache relative Häufigkeitsverteilung
Verteilungsfunktion	Kumulierte relative Häufigkeitsverteilung
Erwartungswert	Arithmetisches Mittel
Varianz	Varianz

    - **Diskrete Zufallsvariable** (diskontinuierliches Merkmal)
      Ein quantitatives Merkmal, das abzählbar (unendlich aber nicht mit unendlichen Zwischenwerten) viele Werte annehmen kann. z.B. Anzahl Mitarbeiter (ein halber Mitarbeiter ist nicht möglich), Anzahl Stück (ein halbes Stück ist nicht möglich)
    - **Stetige Zufallsvariable** (kontinuierliches Merkmal)
      Ein quantitatives Merkmal, das überabzählbar (unendlich mit unendlichen Zwischenwerten) viele Werte annehmen kann. z.B. Körpergrösse, 1.8333m (unendlich)
      Daher umfassen stetige Zufallsvariablen die reellen Werte oder ein Teilbereich davon. Im Beispiel sind es die positiven reellen Werte (Die Rundung auf 5 Rappen wird ignoriert).
      - Die **stetige ZV** kann beschrieben/visualisiert werden durch:
        - ein Histogramm vieler Realisierungen (Näherung) (sh. Abb. 2 **blau**)
        - ihre Dichte (= „geglättetes“ idealisiertes Histogramm, welche das exakte abstrakte Modell repräsentiert) (sh. Abb. 2 **rosarot**)
  - Abb. 2 zeigt die Abb. 1 als **Histogramm.** Dabei ist die Höhe der Säulen durch die Häufigkeitsdichte ersetzt worden. Zudem werden jetzt mehr Klassen abgebildet.
  - Die rosarote **Histogramm-Kurve** stellt die **Dichtefunktion f dar.** Also ein geglättetes idealisiertes Histogramm, welches das exakte abstrakte Modell repräsentiert.
    - Damit die Gesetze der Wahrscheinlichkeitsrechnung erfüllt sind:
      - müssen alle Kurvenwerte ≥ 0 sein
      - muss die Fläche unter der Kurve (Histogramm) = 1 sein.
    - **Die Dichte repräsentiert nun die Wahrscheinlichkeit.**
      Die Ursache liegt in der hohen Durchführung. Wirft man beispielsweise 12'000 Mal einen Würfel, würde jede Zahl ca. 2000 Mal erscheinen und die Wahrscheinlichkeit würde 1/6 betragen.

Abb. 2



- **1.2 (Kumulative) Verteilungsfunktion F**
  - Abb. 3 zeigt nochmals die Dichtefunktion f.
  - Gewisse Aspekte der Zufallsvariablen lassen sich besser durch die **(kumulative) Verteilungsfunktion F** beschreiben (Abb. 4).
  - Die kumulative Verteilungsfunktion ist nichts anderes als die empirische Verteilungsfunktion, nämlich das **Summenpolygon (Ogive),** welche die kumulierten Häufigkeiten F_i angibt (Anteil der Merkmalsträger mit einem Merkmalswert x der kleiner oder gleich x ist). Allerdings stellt F_i nun die Wahrscheinlichkeit dar.
    - Wert von F bei x entspricht der Wahrscheinlichkeit, einen Wert von höchstens x zu beobachten (= Fläche unter der Dichtefunktion f links von x)
    - Beispiel: Mit einer Wahrscheinlichkeit von 80 % haben die Kunden Rechnungen kleiner als 60 Franken.
    - **Eigenschaften von F:**
      - F kann nur Werte zwischen 0 und 1 annehmen
      - F ist monoton steigend
      - F(x) ist das Integral ∫ von f(x)

Abb. 3 **Dichtefunktion f (idealisiertes Histogramm)**

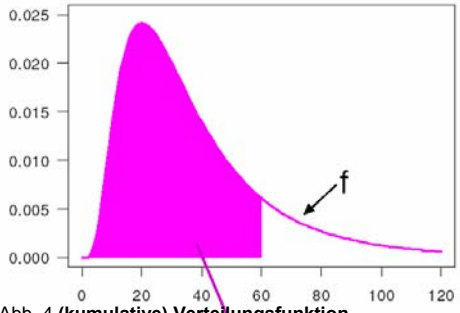
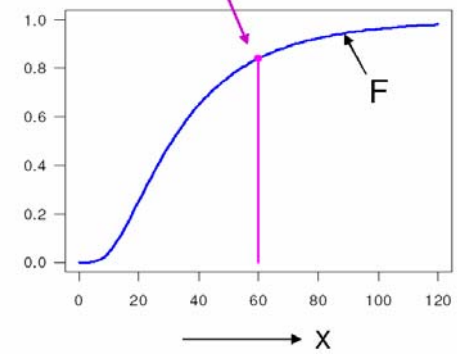


Abb. 4 **(kumulative) Verteilungsfunktion**



- **1.4 Quantile α**
  - Abb. 5 zeigt nochmals die kumulative Verteilungsfunktion F. Die **schwarze** F wiederspiegelt die effektiven empirischen Daten. Die rosarote F zeigt wiederum eine idealisierte Verteilungsfunktion.
    - Beim **Quartil** wird die Gesamtheit n in **vier Viertel** zerlegt.
    - Beim **Median** wird die Gesamtheit n in **zwei Hälften** zerlegt. (bekanntestes Quantil)
    - Beim **Perzentile** wird die Gesamtheit n in **hundert Hundertstel** zerlegt.
    - Beim **Dezentile** wird die Gesamtheit n in **zehn Zehntel** zerlegt.
  - Beispiel:
    - 0.3-Quantil = 30 % Perzentil
    - Mit einer Wahrscheinlichkeit von 30 % kaufen die Leute für Fr. 21.00 oder weniger ein.
    - 30 % kaufen für weniger als Fr. 21.00 ein, 70 % für mehr als Fr. 21.00.
    - **Wäre auch die zugehörige Dichtefunktion f abgebildet, so würde die Fläche links von 21 genau 30 % ausmachen, rechts von 21 70 %.**

Abb. 5

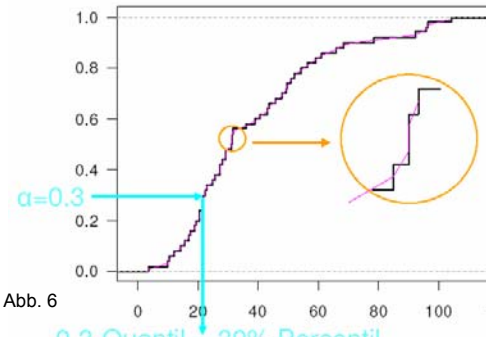
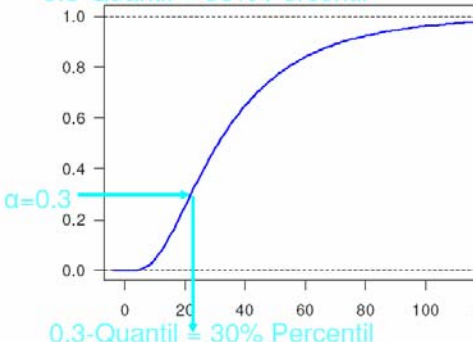


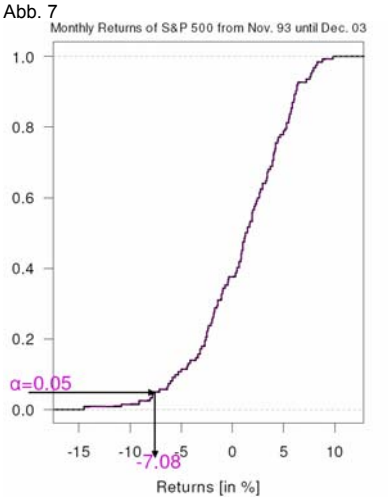
Abb. 6



- **1.5 Value-at-Risk (VaR)** (Anwendung der Quantile)
  - Im Risk-Management werden verschiedene Risikotypen betrachtet. Ein wichtiger Typ darunter ist das **Marktrisiko.** Eines der Masse, um dieses Risiko zu quantifizieren, ist **Value-at-Risk (VaR).**
  - Bei einer einzelnen Anlage wird der VaR oder aus **historischen Daten geschätzt,** indem die Verteilung der Returns betrachtet wird. Dabei wird angenommen, dass die Returns stationär sind, d.h. zukünftige Returns verhalten sich gleich wie vergangene.
  - Diese Schätzung kann im Wesentlichen durch zwei Ansätze erfolgen:
    - Durch eine so genannte **nicht-parametrische** Schätzung
    - Unter der Annahme, dass die Returns einer Modellverteilung folgen, wie z.B. der Normalverteilung
  - **Beispiel**
    - Jemand hält eine $20'000 Position in einem S&P 500 Indexfund. Also entsprechen die Returns jenen des Indexes.
    - Um den VaR zu bestimmen, muss a) **der Zeithorizont** der Anlage und b) **das Konfidenzniveau** festgelegt werden.

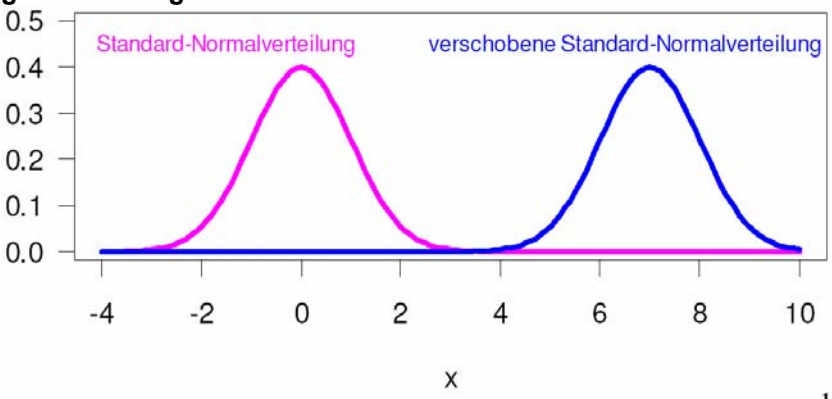
        - Das **Konfidenzniveau** des Anlegers sei 95 %. Das Konfidenzniveau ist ein Vertrauensniveau. Mit einer Wahrscheinlichkeit von 95 % soll der Verlust kleiner sein als ein bestimmter Betrag. In der Regel ist das Konfidenzniveau grösser als 95 %, nämlich so um die 99.9 %!
        - Der **Horizont** betrage bei diesem Fond 1 Monat. Der Horizont ist jene Zeitperiode in der ein Betrag aufgrund von Rückzahlungsbedingungen nicht zurückgezogen werden kann.
    - **Der VaR ist also eine Schranke, sodass der Verlust im Zeithorizont mit Wahrscheinlichkeit 95 % kleiner als diese Schranke ist.**
    - Abb. 7 zeigt die Monatlichen Returns des S&P 500 von Nov. 93 bis Dec. 03.
        - Da das Konfidenzniveau 95 % beträgt, berechnet man das 5 % Quantil aus den monatlichen Returns. Dies ergibt **-7.08 %.**
        - 7.08 % von $20'000 ist $1416
          VaR (1 Monat, 95%) = $1416
          Der Verlust im Zeithorizont von einem Monat ist mit einer Wahrscheinlichkeit von 95 % kleiner als $ 1456. Diesen Betrag würde der Anleger nun absichern (z.B. liquide halten).
        - **Bemerkungen:**
            - Da VaR ein Verlust ist, nehmen wir einen Vorzeichenwechsel im Quantilwert vor.
            - Wenn wir zu wenige Beobachtungen haben, können wir das 5 %-Quantil nicht bestimmen. Bei wenigen Werten ist die statistische Schätzgenauigkeit gering. Alternative: Verwenden von Modellverteilungen.

Abb. 7



- **EXCEL: Histogramm**
    - 3 Spalten
        - 1. Spalte: Klassenuntergrenzen eingeben „0“
        - 2. Spalte: Klassenobergrenzen eingeben „1“
        - 3. Spalte: Absolute einfache Häufigkeit h (leer lassen)
    - Zielbereich gesamte 3. Spalte markieren
    - Einfügen > Funktion > Statistik > Häufigkeit
        - Daten eingeben: Gemäss Aufgabenstellung jene Daten die in Klassen zu unterteilen sind
        - Klassen eingeben: 2. Spalte
        - Fenster nicht mit „Enter“ schliessen, sondern mit „Shift-Control-Enter“
    - 4. Spalte: Relative einfache Häufigkeit f (h-Werte durch Total, wobei Total mit F4 unveränderbar setzen)
    - Einfügen > Diagramm > Gruppierte Säulen
        - Datenbereich: 4. Spalte
        - zu Reihe umschalten
        - Beschriftung der Rubrikenachse: 1. und 2. Spalte
        - Diagramm beschriften (Y-Achse: rel. einfache Häufigkeit, X-Achse: Klassen, Titel: Histogramm)
        - Diagramm abschliessen
        - Säulen anklicken > Optionen > Abstandsbreite: 0 und Punktfarbunterscheidung
- **EXCEL: Dichtefunktion f**
    - Gleiches Vorgehen wie Histogramm, aber:
        - Diagrammtyp: Linie
        - keine Punktfarbunterscheidung
        - Neuer Titel: Dichtefunktion f
- **EXCEL: Verteilungsfunktion F**
    - Gleiches Vorgehen wie Histogramm und Dichtefunktion, aber
        - 5. Spalte: F Relative kumulierte Häufigkeit
        - Diagrammtyp: wie Dichtefunktion
        - Datenquelle: 5. Spalte
- **EXCEL: Quantile**
    - Verteilungsfunktion F erstellen
    - Einfügen > Funktion > Statistik > Quantil
        - Matrix: Ursprungsdaten auswählen gemäss Aufgabenstellung (nicht die 3. Spalte)
        - Alpha: Prozentwert von 0.0 bis 1
- **EXCEL: VaR**
    - Quantil suchen wie oben beschrieben (100 % - Konfidenzniveau)
    - Betrag in $ Mal %-Satz

2. Normal- oder Gauss-Verteilung, Erwartungswert, Varianz

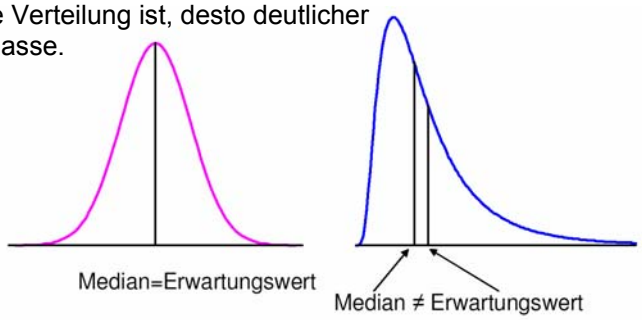
- **2.1 Normal- oder Gauss-Verteilung**
  - Das **üblichste stetige Verteilungsmodell** ist die **Normal- oder Gaussverteilung.** Sie spielt auch in der Finanztheorie eine zentrale Rolle („Black-Scholes-Theorie").
  - Über ein Histogramm wird eine Idealisierung gelegt: Diese Idealisierung erhält eine **Glockenform.** Diese **Normalverteilung** beschreibt also eine **Zufallsvariable** mit einer **glockenförmigen Dichte.**
  - 

    - **Standardnormalverteilung,** Funktionsgleichung: $\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$
    - **(Verschobene) Normalverteilung,** Funktionsgleichung: $f(x) = \frac{1}{\sqrt{2 \cdot \pi \cdot \sigma^2}} \cdot e^{-(x-\mu)^2 / 2\sigma^2}$

- **2.2 Der Erwartungswert** (Lageparameter)
  - Eine <u>erste Erweiterung</u> der Standard-Verteilung erreicht man, indem horizontale Verschiebungen zugelasen werden, die sogenannten **Lageparameter**:
    1. Durch **Median;** d.h. die halbe Fläche unter der Dichtekurve liegt links vom Median, die andere Hälfte rechts davon. Weil die Normalverteilung symmetrisch ist, fällt der Median mit dem **Symmetriezentrum** zusammen.
    2. Durch den **Erwartungswert E(X),** der eine **Idealisierung des arithmetischen Mittels** ist. Wie das arithmetische Mittel, lässt sich auch der Erwartungswert als **Massenschwerpunkt** interpretieren: Es ist der Punkt bei welchen die Dichtekurve im Gleichgewicht wäre, wenn sie aus festem Material bestehen würde.

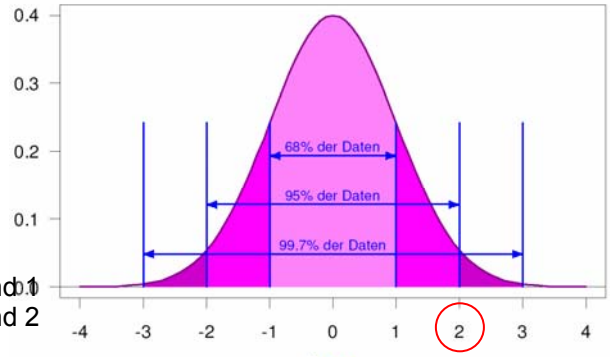
    

    - Der **Erwartungswert E(X)** der Zufallsvariablen X lässt sich berechnen mit: $E(X) = \int_{-\infty}^{\infty} x \cdot f(x) dx$ Dabei ist **f(x)** die Dichte von X.
    - Der Erwartungswert wird mit dem griechischen Buchstaben **μ (mü, m = Mitte)** gekennzeichnet.
    - Im obigen Beispiel hat die **Standard-Normalverteilung** ihr Zentrum bei x = 0, daher **μ (mü) = 0.** Die **(Verschobene) Normalverteilung** hat ihr Zentrum bei x = 7, daher **μ (mü) = 7**
  - Bei **symmetrischen Verteilungen** wie der Normalverteilung **kennzeichnen diese beiden Lagemasse denselben Punkt.** Je **schiefer** die Verteilung ist, desto deutlicher **unterscheiden sich aber diese** beiden Lagemasse.
    - Bei symmetrischer Verteilung (Normalvertielung): **Median = Erwartungswert** (sh. Abbildung)
    - Bei rechtschiefer Verteilung (keine Normalverteilung): **Median < Erwartungswert** (sh. Abbildung)
    - Bei linksschiefer Verteilung (keine Normalverteilung): **Median > Erwartungswert**

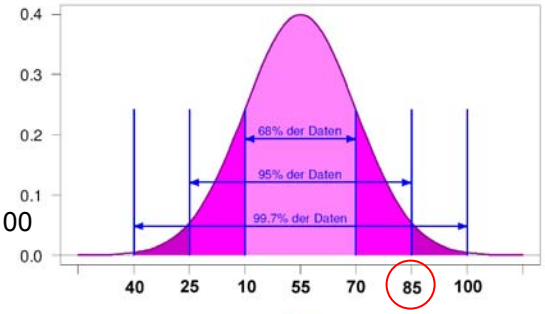
    - Wie gesagt ist die Ursache hiefür, dass rechts und links vom Median die gleiche Fläche ist, beim Erwartungswert handelt es sich aber um den Massenschwerpunkt.

- **2.3 Varianz und Standardabweichung** (Streumasse)
  - Eine zweite Erweiterung der Standard-Normalverteilung erreicht man, indem die Weite der Glockenkurve variiert wird. Diese Weite kann mit dem so genannten **Streu- oder Skalenparameter** kontrolliert werden. Die Streuung kann gemessen werden:
    1. Durch den **Zentralen Quartilsabstand (ZQA)** (Interquartile-Range) (3. Quartil – 1. Quartil)
    2. Durch die idealisierte **Standardabweichung:**
       - **Varianz σ^2 :** $\sigma^2 = \int_{-\infty}^{\infty} (x - E(X))^2 \cdot f(x)\,dx$
       - **Standardabweichung σ :** $\sqrt{\sigma^2}$
  - **Alle Normalverteilungen** lassen sich **gleich charaktisieren**, falls die Realisierungen in Einheiten von σ (Standardabweichung) um den Erwartungswert μ (hier = 0) beschrieben werden.
    - **Standardnormalverteilung**
      - Daher:
        68 % der Daten liegen zwischen -1 und 1
        95 % der Daten liegen zwischen -2 und 2
        99.7 % der Daten liegen zw. -3 und 3



in σ **Z** bei Standard-NV

    - **Verschobene Normalverteilung**
      - Beträgt der Erwartungswert μ = 55 und die Standardabweichung σ = 15 dann gilt in der nebenstehenden Darstellung folgendes:
      - Daher:
        68 % der Daten liegen zwischen 40 und 70
        95 % der Daten liegen zwischen 25 und 85
        99.7 % der Daten liegen zwischen 10 und 100



in σ **X** bei verschob. NV

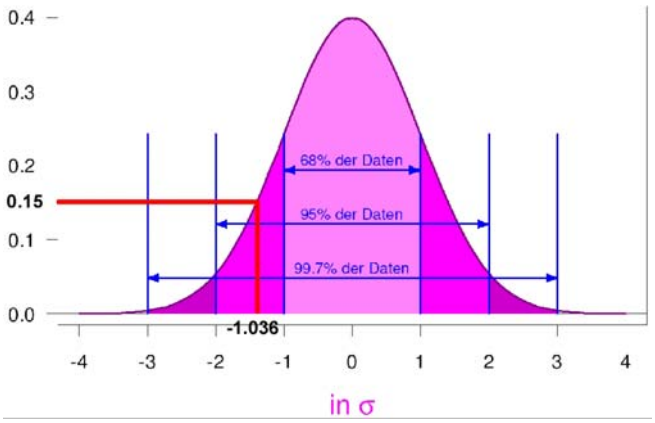
2.4 Kurzschreibweisen

- Eine **Zufallsvariable** (Vorgang Würfelwurf) bezeichnen wir mit einem **Grossbuchstaben;** oft mit X, Y und Z.
  Eine **Realisierung** (Resultat Würfelzahl) bezeichnen wir mit einem kleinen Buchstaben; oft mit x, y und z.
- Dass die **Zufallsvariable X normalverteilt** (nicht zwingend Standard-Normalverteilung sondern verschobene Normalverteilung) mit **Erwartungswert μ** und **Standardabweichung σ** ist, wird oft durch die Kurzschreibweise **X ist N(μ, σ)** festgehalten.
- **Verschiebung der Normalverteilung (z-Transformation)**
  - Jede verschobene Normalverteilung N(μ, σ) kann in die Standard-Normalverteilung N(0,1) transferiert werden. Hiefür ist folgende Formel notwendig:
  - Man will vom **X-Wert** (irgend ein Wert der verschobenen Normalverteilung) **der verschobenen Normalverteilung** auf den **Z-Wert der Standard-Normalverteilung** schliessen.
  - $Z = \frac{X - \mu}{\sigma}$ oder **X = μ + σ · Z** (Zurückrechnen Standard → Verschoben)
  - **Beispiel:**
    - Verschobene Normalverteilung (sh. Beispiel 2.3): X = 85
      Umrechnung: Z = (85 – 55) / 15 = 2
      Standard-Normalverteilung: Y = 2

- **2.5 Quantile bei einer Normalverteilung**
  - Die Quantile einer verschobenen normalverteilten Zufallsvariablen X lassen sich aus der Standnormalverteilung berechnen: **α-Quantil von N(μ, σ) = μ + σ · (α-Quantil von N(0,1)**
  - Die Quantile der Standardnormalverteilung müssen in Tabellen (sh. unten) nachgeschlagen werden oder mit Excel bestimmt werden, da sie nicht explizit berechnet werden können.
  - **Beispiel: Quantil**
    - Das 0.15-Quantil von N(μ=5, σ=3):
      5 + 3 · (-1.036) = 1.8922 [Excel: =STANDNORMINV(0.15) = -1.036]
    - Die Wahrscheinlichkeit, dass eine Beobachtung unter 1.8922 liegt, beträgt 15 %.
    - Die Wahrscheinlichkeit, dass eine Beobachtung über 1.8922 liegt, beträgt 85 %

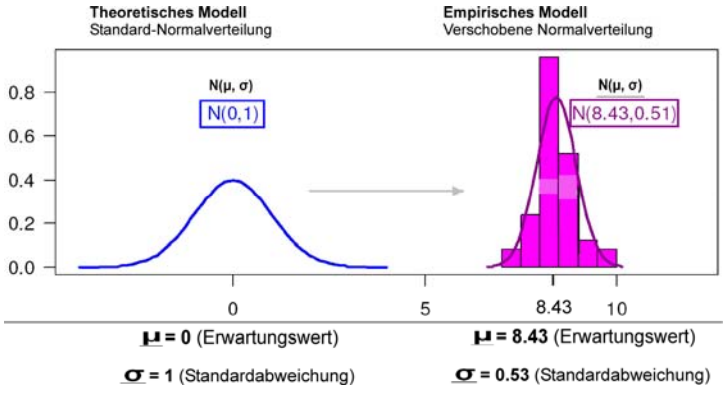
- **Beispiel 1: VaR**
  - Nehmen wir an, wir haben eine Anlage von Fr. 1'000'000.00. Die Anlage liefert jährliche Returns, die normalverteilt sind mit Erwartungswert 6.45 % und Standardabweichung 3.75 %. Der VaR bei einem Zeithorizont von 1 Jahr und einem Konfidenzniveau von 0.99 ist:
  - [6.45 + 3.75 · (-2.326 $_{(0.01\text{-Quantil der Standardnormalverteilung})}$)] = -2.2725 / **100** · 1'000'000 = 22'725.00
    oder: =STANDNORMINV(0.01) = -2.326
  - [0.0645 + 0.0375 · (-2.326 $_{(0.01\text{-Quantil der Standardnormalverteilung})}$)] = -0.022725 · 1'000'000 = 22'725.00
  - Mit einer Wahrscheinlichkeit von 99 % wird mein Verlust kleiner als 22.725 sein.
- **Quantile der Standard-Normalverteilung**

0.50	±0.000	**0.75**	±0.674
0.51	±0.025	**0.76**	±0.706
0.52	±0.050	**0.77**	±0.739
0.53	±0.075	**0.78**	±0.772
0.54	±0.100	**0.79**	±0.806
0.55	±0.126	**0.80**	±0.842
0.56	±0.151	**0.81**	±0.878
0.57	±0.176	**0.82**	±0.915
0.58	±0.202	**0.83**	±0.954
0.59	±0.228	**0.84**	±0.994
0.60	±0.253	**0.85**	±1.036
0.61	±0.279	**0.86**	±1.080
0.62	±0.306	**0.87**	±1.126
0.63	±0.332	**0.88**	±1.175
0.64	±0.358	**0.89**	±1.226
0.65	±0.385	**0.90**	±1.282
0.66	±0.412	**0.91**	±1.341
0.67	±0.440	**0.92**	±1.405
0.68	±0.468	**0.93**	±1.476
0.69	±0.496	**0.94**	±1.555
0.70	±0.524	**0.95**	v1.645
0.71	±0.553	**0.96**	±1.751
0.72	±0.583	**0.97**	±1.881
0.73	±0.613	**0.98**	±2.054
0.74	±0.643	**0.99**	±2.326



- **Beispiel:** Gesucht ist das 0.15-Quantil
  - 1 – 0.15 = **0.85** -1.036

3. Parameterschätzung, Quantil-Quantil-Plot



- **3.1 Parameterschätzung**
  - **Was ist die Parameterschätzung?**
    - Das Ziel der **Parameter-schätzung** ist es, die Daten aus einem Histogramm mit einer Normalverteilung zu beschreiben (Wir sagen dem auch **modellieren.**).
    - Mit der Parameterschätzung werden die beiden Parameter μ (Erwartungswert) und σ (Standardabweichung) so festgelegt, dass die entsprechende Normalverteilung möglichst gut passend über die Daten zu liegen kommt. Dies nennen wir **„Schätzen von μ und σ".**
    - Nicht in allen Fällen lassen sich jedoch die betrachteten Daten genügend gut durch die Normalverteilung beschreiben (siehe 3.2 Eignet sich die Normalverteilung?).
  - **Wie führen wir die Parameterschätzung durch?**
    - Die beiden Parameter μ und σ werden dadurch **geschätzt,** dass **μ** mit seinem Synonym aus der beschreibenden Statistik, dem arithmetischen Mittel und **σ** auch mit seinem Synonym aus der beschreibenden Statistik, der Stichproben-Standardabweichung, gleich gesetzt wird.

$$\mu = \text{Arithmetisches Mittel } (\bar{x}) = \frac{\sum_{i=1}^{k} (x'_i \cdot h_i)}{n}$$

x'_j = Klassenmitte =MITTELWERT()

$$\sigma = \text{Standardabweichung (s)} = \sqrt{\frac{1}{n-1} \cdot \sum_{i=1}^{k} [(x_i - \bar{x})^2]}$$

x_i = Klassenmitte =STABW()

    - Beispiel: In der letzten Abbildung war das arithmetische Mittel der Daten, die im Histogramm dargestellt sind, 8.43 und die Stichproben-Standardabweichung war 0.51.

- Es bleiben nun noch **zwei gewichte Fragen** offen:
  - Sind die Daten überhaupt **normal verteilt?** (sh. **3.2**)
  - Die Schätzer werden je nach ausgewählten Daten (andere Realisierungen desselben Modells) leicht andere Resultate liefern. Wenn ich also nochmals Daten erheben würde, wird das arithmetische Mittel immer gleich sein? Wie gross sind die Unterschiede? (sh. **3.3**)

- **3.2 Eignet sich die Normalverteilung?**
  - Die Entscheidung, die Daten durch die **Normalverteilung zu modellieren,** steht am Anfang einer statistischen Analyse und beeinflusst grundsätzlich das weitere Vorgehen. Deshalb ist diese **Entscheidung** folgenschwer und sollte entsprechend wohlüberlegt gemacht werden.
  - Die **Entscheidung beruht** meistens auf den **folgenden zwei Grundpfeilern:**

1. Theoretische Überlegungen

- Theoretische Überlegungen aus dem vorliegenden Fachgebiet (z.B. aus der Finanztheorie) können fordern, dass die betrachteten Daten normalverteilt sein müssten.
- Beispiel: Sei P_t der Preis einer Anlage zur Zeit t und $R_r = (P_t - P_{t-1}) / P_{t-1}$ der Return zur Zeit t.
  - Modell: R_t ist $N(\mu, \sigma)$
    Damit gibt es aber zwei Probleme:
    1. Eine normalverteilte Zufallsvariable kann Werte zwischen $-\infty$ und $+\infty$ annehmen (Normalverteilungs-Kurve läuft unendlich nach rechts und links), aber R_t kann oft nicht tiefer als -1 sein (nämlich dann, wenn $P_t = 0$ ist).
    2. Mehrperioden-Returns sind nicht mehr normalverteilt, wenn Einperioden-returns R_t normalverteilt sind (Fakt aus der Wahrscheinlichkeitstheorie)
  - **Diese Defizite können behoben werden, falls die Log-Returns $r_t = \log(P_t / P_{t-1})$ als normalverteilt werden.**

2. Explorative Datenanalyse

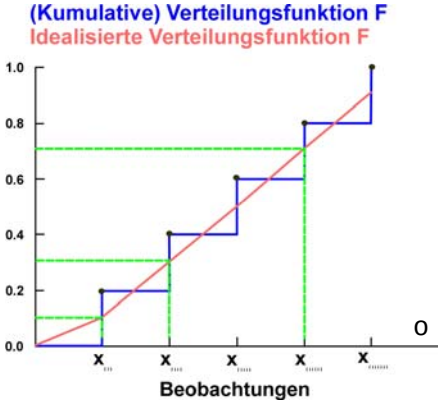
- Die explorative Datenanalyse sagt aus, dass aus einem **Histogramm** oder aus einem **Quantil-Quantil-Plot (Q-Q-Plot) erkennbar ist,** ob die Daten durch eine **Normalverteilung beschreibbar sind oder nicht.**
  - **Histogramm**
    - Das Histogramm kann ausgeprägte Abweichungen von einer Normalverteilung wie Schiefe, Ausreisser und Lücken aufzeigen.
    - Die Einschätzung kann mit der **68-95-99.7 %-Regel** schlagkräftiger gemacht werden. **Daher, wir zählen zusätzlich aus, wie viele Beobachtungen innerhalb von $\bar{x} \pm s$, $\bar{x} \pm 2s$, $\bar{x} \pm 3s$, liegen. s = Standardabweichung, $\bar{x}$ = Arithmetisches Mittel**
  - **Quantil-Quantil-Plot (Q-Q-Plot)**
    - Durch den Q-Q-Plot kann eine Überprüfung vorgenommen werden, die oft noch deutlicher aufzeigt, ob eine Normalverteilung vorliegt.
    - **Schritt 1: Bestimmung geordnete Beobachtungen**
      - Zuerst werden die **geordneten Datenwerte** bestimmt.
      - Die geordneten Datenwerten entsprechen dabei folgenden Quantilen:

$\left(\frac{1}{n} - \frac{1}{2n}\right)$ **- Quantil = Kleinste Beobachtung $x_{(i)}$**

$\left(\frac{2}{n} - \frac{1}{2n}\right)$ **- Quantil = Zweit-kleinste Beobachtung $x_{(ii)}$**

$\left(\frac{n}{n} - \frac{1}{2n}\right)$ **- Quantil = Grösste Beobachtung $x_{(n)}$**

n entspricht dabei der Gesamtzahl der Merkmalsträger.

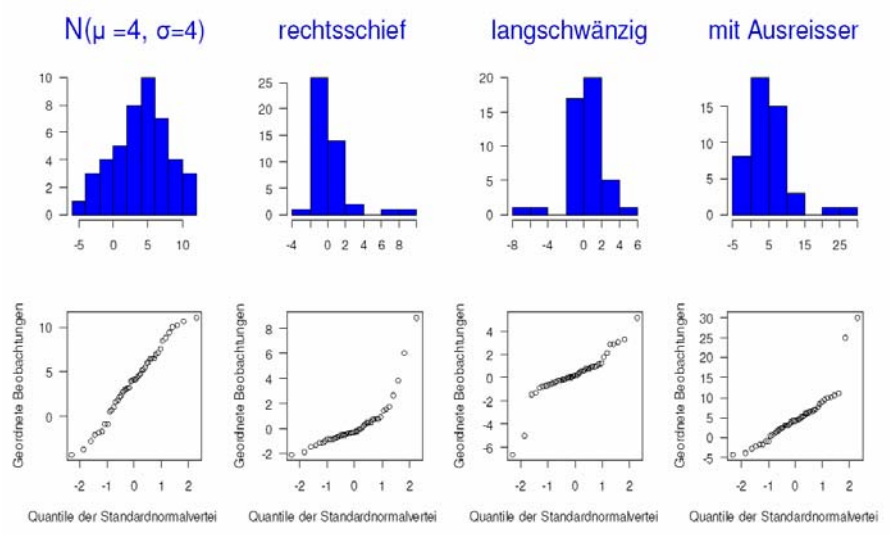


      - **Beispiel:** Die linksstehende Grafik zeigt eine kumulative Verteilungsfunktion. Angenommen die kumulierten Beobachtungen betragen $x_{(i)} = 1$, $x_{(ii)} = 2$, $x_{(iii)} = 3$, $x_{(iiii)} = 4$, $x_{(iiiii)} = 5$.
        $\left(\frac{1}{5} - \frac{1}{10}\right)$ -Quantil entspricht dem 0.1-Quantil. Das 0.1-Quantil in dieser Verteilungsfunktion beträgt gerade $x_{(i)} = 1$ wie vorgehend definiert. **Das heisst, mithilfe der obigen Formeln zur Bestimmung der bestimmten Quantile kann man gerade die genaue Beobachtung bestimmen, also eine geordnete Beobachtung.**
    - **2. Schritt: Quantile der Standardnormalverteilung**
      - In einem zweiten Schritt werden die gleichen Quantile, z.B. $\left(\frac{1}{n} - \frac{1}{2n}\right)$ wie oben jetzt bei der Standardnormalverteilung bestimmt.

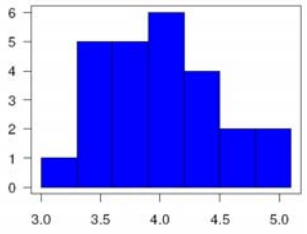
- **3. Schritt: Q-Q-Plot erstellen**
  - Der Q-Q-Plot wird folgendermassen erstellt:

    $$\left[x(i);\ \left(\frac{1}{n} - \frac{1}{2n}\right)\text{-Quantil von } N(0,1)\right]$$

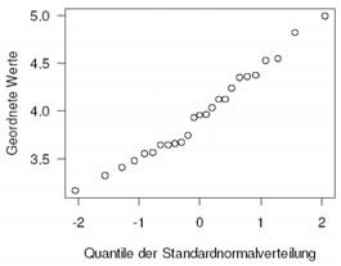
  - Das gleiche für alle anderen Punkte machen x(ii), x(iii) bis x(n).
- **Q-Q-Plot Beispiele:**
  - In der folgenden Darstellung sind 4 verschiedene Q-Q-Plots dargestellt.



  - **Falls die Daten genügend gut durch eine Normalverteilung modelliert werden können, streuen die Dateien im Q-Q-Plot um eine Gerade mit Achsenabschnitt μ und Steigung σ.**
  - In obigen Beispielen zeigt dies deshalb, dass nur dem ersten Q-Q-Plot Daten zugrunde liegen, die durch eine Normalverteilung beschrieben werden können. Die restlichen drei Histogramme sind nicht normalverteilt.



- **3.3 Schätzen ist mit Unsicherheit behaftet.**
  - Die Schätzer werden je nach ausgewählter Stichprobe (andere Realisierungen desselben Modells) leicht andere Resultate liefern.
  - **Beispiel:**
    - 45 Experimente wurden durchgeführt. Die daraus ermittelten arithmetischen Mittelwerte wurden in einem Histogramm und dem Q-Q-Plot aufgezeichnet. Einmal ist das arithmetische Mittel 3.163, beim nächsten Versuch dann 3.326, dann 4.363 und dann wieder 4.120…
    - **Kann man die Schätzungen mit neuen Stichproben wiederholen, so sieht man, dass die Schätzwerte streuen wie eine Normalverteilung.**



4. Der statistische Hypothesentest (Teil 1)

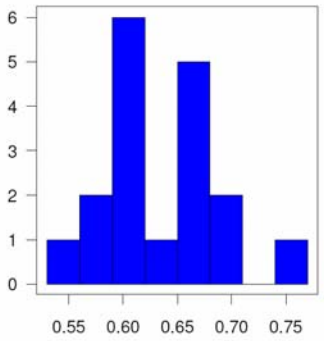
- **4.1 Beispiel: Rechtecke in der Kunst**
  - Schon in der Antike hatte die Gesellschaft Vorlieben für bestimmte Verhältnisse zwischen Breite und Länge bei Rechtecken. Der **Goldene Schnitt** galt schon bei den Alten Griechen als ein „ästhetischer Standard“; d.h. das Verhältnis von Breite zu Länge ist gleich **0.618**.

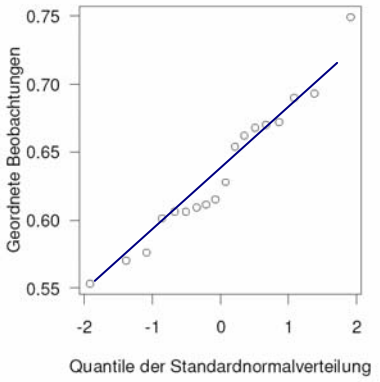
    $$\frac{\text{Breite}}{\text{Länge}} = 0.618$$

  - Dieses ästhetische Grundgesetz findet sich seither auch immer wieder in der Architektur und im Design (z.B. Visitenkarten, Kreditkarten, ID, Bilderrahmen).
  - **Man kann sich fragen, ob die Shoshoni-Indianer,** welche keine Verbindungen zum antiken Griechenland hatten, **den Goldenen Schnitt ebenfalls als einen ästhetischen Standard betrachteten?** Die folgenden Verhältnisse wurden von **18 Rechtecken** auf verschiedenen Kunsthandwerkarbeiten (perlenbestickte Rechtecke) der Shoshoni-Indianer gewonnen:

0.693	0.749	0.654	0.670	0.662	0.672	0.615	0.606	0.690
0.628	0.668	0.611	0.606	0.609	0.601	0.553	0.570	0.576

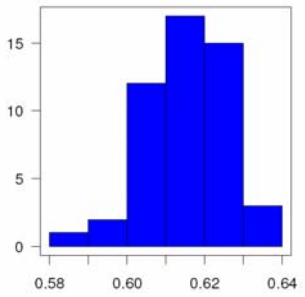
- Hilfreich wäre es, wenn die Verhältnisse **normalverteilt** sind. Also betrachten wir das Histogramm und den Q-Q-Plot:
- Die Daten scheinen weder langschwänzig noch mit groben Ausreisser durchsetzt zu sein. Die sichtbaren Strukturen können Effekte der kleinen Zahl der Beobachtungen sein. Es liegt also eine **Normalverteilung vor.**

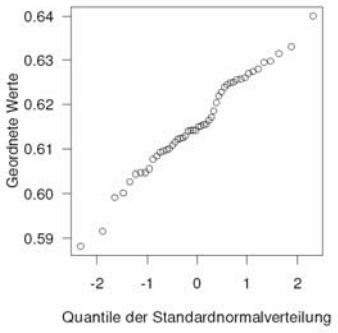




- Also schätzen wir den **Erwartungswert** und die **Standardabweichung der Daten:**
  - Erwartungswert (Arithmetisches Mittel): $\bar{x}$ = 0.635
  - Standardabweichung: s = 0.050
- **Offensichtlich ist aber $\bar{x}$ = 0.635 nicht identisch zum Goldenen Schnitt (φ = 0.618). Es stellt sich deshalb die Frage, ob dieser Unterschied nur auf <u>Ungenauigkeiten bei der Erhebung der Daten</u> zurückzuführen ist oder ob wirklich ein Unterschied besteht!**

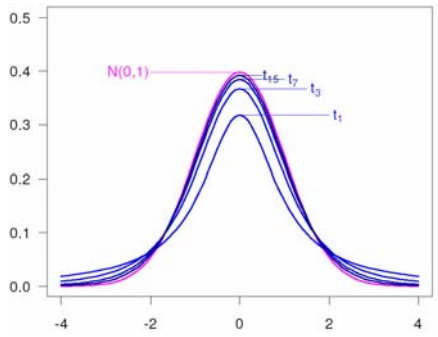
- **4.2 Was wäre, wenn die Hypothese richtig wäre?**
  - **Wie müsste sich das arithmetische Mittel verhalten, wenn die Verhältnisse tatsächlich dem Goldenen Schnitt (φ = 0.618) entsprächen?**
  - Im Kapitel 3.3 haben wir festgestellt, dass **Schätzen mit Unsicherheit** verbunden ist und dass jedes Experiment (andere 18 Datensätze der Shoshoni-Indianer) ein anderes arithmetisches Mittel ergibt.
  - Nehmen wir an, dass die Verhältnisse normalverteilt seien mit μ = 0.618 und σ = 0.050.
  - Anschliessend können wir zusätzliche **50 Experimente (Stichproben)** durchführen, daher **nochmals 18 Realisierungen ziehen,** Mittelwert bilden, nochmals 18 Realisierungen ziehen, Mittelwert bilden, usw. bis wir z.B. 50 Mittelwerte haben. Die **50 Mittelwerte** bilden dann den nebenstehenden Q-Q-Plot.



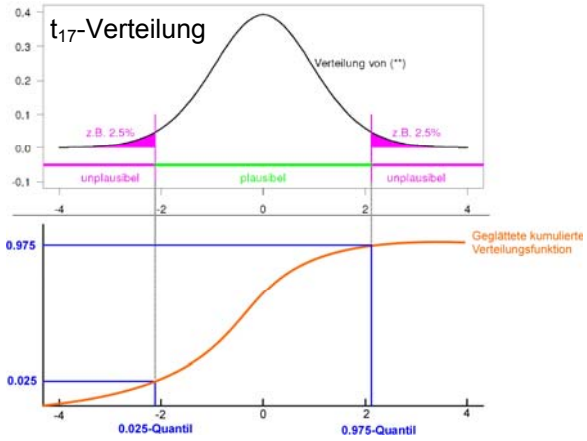


  - Diese 50 arithmetischen Mittel aus dem Experiment scheinen auch **normalverteilt** zu sein.
  - Also schätzen wir den **Erwartungswert** und die **Standardabweichung der Daten:**
    - Erwartungswert (Arithmetisches Mittel): $\bar{x}$ = 0.6156
    - Standardabweichung: s = 0.0106
  - **Offensichtlich ist aber $\bar{x}$ = 0.6156 auch bei mehr als einem Experiment nicht identisch zum Goldenen Schnitt (φ = 0.618). Doch 0.6156 ist schon ziemlich nahe.** Wenn wir mehrere Experimente durchführen wird sich der arithmetische Mittelwert also bereits dem goldenen Schnitt annähern. Das war beim einen Experiment mit Mittelwert von 0.635 nicht der Fall.
  - <u>Erkenntnis:</u>
    Stammen die n Datenwerte eines Experiments (hier n = 18 Datenwerte) **von einer N(μ, σ)-Verteilung** (das ist vorliegend erfüllt, sh. Experiment 1 bei Kap. 4.2 denn N(0.635, 0.050))**, dann ist das arithmetische Mittel aller arithmetischen Mittel der Experimente** (Arithmetisches Mittel aller arithmetischen Mittel der Experimente 1 – 50) **N(μ, σ/√n)-verteilt.**
    - **Also im vorliegenden Fall hatten wir eine N(φ = 0.635, σ = 0.050)-Verteilung in einem ersten Experiment. Die Mittelwerte aller Experimente müssten also folgendermassen verteilt sein: N(μ = 0.618, σ = 0.050/√18).** Das heisst N(μ = 0.618, σ ≈ 0.0118). Das stimmt in etwa überein mit unseren Berechnungen: $\bar{x}$ = 0.6156 und s = 0.0106
    - Das kann man noch umformulieren: $\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$ *ist* $N(0,1)$
  - In unserem Beispiel haben wir nur eine Hypothese über μ (0.618), jedoch nicht bezüglich σ. Ausweg: Wir **bestimmen** die **Standardabweichung aus den Daten** und bilden: $\frac{\bar{x}-\mu}{s/\sqrt{n}}$ (**)

- Dass die **Standardabweichung nicht bekannt** ist und aus den **Daten geschätzt** werden muss, hat aber **Konsequenzen:**
  - **(**) ist nicht mehr N(0,1),** sondern verteilt wie die so genante **t-Verteilung mit (n-1) Freiheitsgrade (t_{n-1})**
  - Vorliegend ist eine t_{18-1}, also eine t_{17}-Verteilung. Je mehr Freiheitsgrade, also Realisationen man hat, desto eher wird die Normalverteilung zu einer Standard-Normalverteilung.



- **4.3 Welche Mittelwerte sind unter der Hypothese plausibel?**
  - Prinzipiell ist jeder Mittelwert möglich, aber nicht alle Werte sind unter der Hypothese gleich plausibel.
  - **Wir deklarieren alle Werte, die in den beiden extremen Enden der Verteilung liegen als unplausibel.** Grafisch sieht das so aus:
  - Wir gehen somit ein Risiko ein, in 5 % (= 2.5 % + 2.5 %) der Fälle (**) als unplausibel zu bezeichnen, obwohl der Wert möglich wäre. Dieses Risiko wird **Signifikanzniveau €** genannt.



Zurück zu unserem Beispiel:
  - Wir schätzten den Erwartungswert und die Standardabeichung aus dem Experiment 1: ⌡ = 0.635 und s = 0.050
  - Wir bilden nun **(**)**: $\frac{\bar{x}-\mu}{s/\sqrt{n}}$ (**) $\frac{0.635-0.618}{0.050/\sqrt{18}} = 1.44250$
  - Wir bestimmen das **Signifikanzniveau:** Es soll hier 5 % sein. Deshalb bestimmen wir das 0.025- und das 0.975-Quantil der t_{17}-Verteilung: -2.1098 und 2.1098 (siehe obere Grafik).
  - **Da unser (**)-Wert (1.44250) innerhalb dieser beiden Grenzen liegt, ist die Hypothese, dass die Shoshoni-Indianer den Goldenen Schnitt ebenfalls als einen ästhetischen Standard betrachteten, für uns plausibel.**
  - Wir sagen auch: Die Hypothese, dass die Shoshoni-Indianer den Goldenen Schnitt ebenfalls als einen ästhetischen Standard betrachten, kann auf dem 5 %-Niveau nicht verworfen werden.

- **4.4 Schema zum Einstichproben-t-Test**
  1. **Hypothesen formulieren:**
     Null-Hypothesen: „Shoshoni-Indianer betrachten den Goldenen Schnitt als einen ästhetischen Standard."
     Alternative: „Shoshoni-Indianer haben einen anderen ästhetischen Standard."
  2. **Prüfgrösse und deren Verteilung unter der Nullhypothese festlegen:**
     Prüfgrösse = $\frac{\bar{x}-0.618}{s/\sqrt{18}}$ 0.618 = μ (optimaler Erwartungswert), √18 = Anzahl Realisationen
     : Sie ist t_{17}-verteilt.
  3. **Signifikanzniveau und daraus plausiblen Bereich festlegen:**
     Signifikanzniveau sei 5 %, also beide Enden der t_{17}-Verteilung mit 2.5 % + 2.5 % (sh. Grafik oben).

     Gesucht ist also das 0.025-Quantil und das 0.975-Quantil der t_9-Verteilung: $[t_9\ q_{0.025};\ t_9\ q_{0.975}]$
     Mit EXCEL kann das 0.975-Quantil von t_{17} wie folgt berechnet werden:
     =TINV(α;17)
     α berechnet sich wie folgt:
     1 – α/2 = 0.975 / -1
     -α/2 = -0.025 /·(-2)
     α = 0.05
     =TINV(0.05;17) = 2.1098
     Für das 0.025-Quantil einfach das Vorzeichen wechseln = -2.1098

     Plausibler Bereich ist also zwischen -2.1098 und 2.1098, → [-2.1098; 2.1098]

4. **Prüfgrösse berechnen:**
   Zuerst Arithmetisches Mittel und Standardabweichung aus den Daten (18 Realisationen) ermitteln:

   Arithmetisches Mittel: $\bar{x} = \frac{\sum_{i=1}^{k} (x'_i \cdot h_i)}{n}$ =MITTELWERT() = 0.635

   Standardabweichung: $(s) = \sqrt{\frac{1}{n-1} \cdot \sum_{i=1}^{k} [(x_i - \bar{x})^2]}$ =STABW() = 0.050

   Prüfgrösse = $\frac{0.635 - 0.618}{0.050/\sqrt{18}} = 1.44250$: Sie ist t_{17}-verteilt. $\frac{\bar{x} - \mu}{s/\sqrt{n}}$ (**)

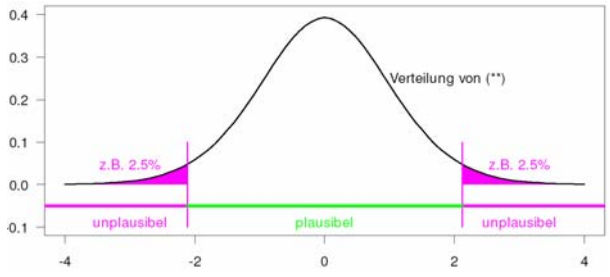
5. **Entscheiden, ob plausibel oder nicht:**
   Da die Prüfgrösse innerhalb des plausiblen Bereichs liegt, kann die Hypothese H_0 auf dem 5 %-Niveau nicht verworfen werden. H_0 ist plausibel.

5. Der statistische Hypothesentest (Teil 2), Vertrauensintervall

- **5.1 Risiko von Fehlentscheiden**
  - Aussagen in Testverfahren kann man wie gesagt wegen des Einflusses des Zufalls (Rauschen) auf das Ergebnis nicht mit Sicherheit machen. In der Methode selber liegt eine Unsicherheit.
  - Es bleibt somit immer ein **Restrisiko,** einen falschen Schluss zu ziehen:

		Unser Entscheid	
		H_0 ist plausibel (H_0 = Nullhypothese)	H_0 ist unplausibel (verwerfe H_0)
Tatsächlich gilt	H_0 ist richtig	Richtiger Entscheid	Falscher Entscheid: **Fehler 1. Art**
	H_0 ist falsch, Alternative ist richtig	Falscher Entscheid: **Fehler 2. Art**	Richtiger Entscheid

  - **Fehler 1. Art** (Unser Entscheid im Testverfahren: H_0 unplausibel / Tatsächlich gilt: H_0 ist richtig)
    - **Die Wahrscheinlichkeit, einen Fehler 1. Art zu begehen, ist ja gerade das Signifikanzniveau α.**
    - Dieses Risiko kontrollieren wir also mit der Wahl des Signifikanzniveaus.
    - Beispiel:
      - Im Beispiel mit den Shoshoni-Indianern wurde das Signifikanzniveau auf 5 % (2.5 % + 2.5 %) festgelegt.
      - Per Zufall kann die Prüfgrösse im unplausiblen Bereich liegen, obwohl sie in Wirklichkeit im plausiblen Bereich liegt. Damit entsteht ein Fehler 1. Art.

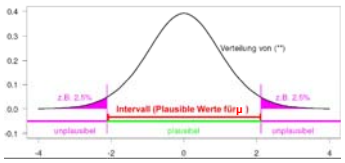


  - **Fehler 2. Art** (Unser Entscheid im Testverfahren: H_0 ist plausibel / Tatsächlich gilt: H_0 ist falsch)
    - Um die **Wahrscheinlichkeit, einen Fehler 2. Art zu begehen,** zu bestimmen, müssten wir mehr über die Alternative wissen. Üblicherweise werden wir aber zu wenig wissen, um Rechnungen darüber durchführen zu können.
    - Beispiel:
      - Eine Alternative wäre Beispiel ein anderes Ideal für einen goldenen Schnitt bei den Shoshoni-Indianern. Jedoch kennen wir kein solches und können es daher auch nicht überprüfen.
    - Folglich werden **wir nie wissen, welches Risiko wir eingegangen sind,** wenn wir die Null-Hypothese als plausibel erachten.
    - Beispiel**:**
      - Im Beispiel mit den Shoshoni-Indianern wissen wir also nicht, mit welchem Risiko (=Wahrscheinlichkeit) wir eine Fehlentscheidung (Fehler 2. Art) getroffen haben. Das ist natürlich sehr unbefriedigend, aber wir können dieses Problem nicht umgehen.
    - **Eine Nullhypothese kann nie statistisch bewiesen werden. Daten können nur gegen eine solche sprechen.**

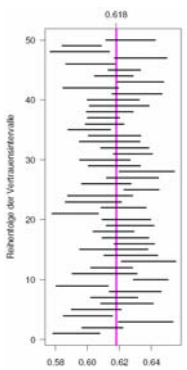
- **Warum machen wir das Risiko, einen Fehler 1. Art zu begehen (= Signifikanzniveau α) nicht beliebig klein?**
  - Indem wir das Signifikanzniveau verkleinern, wird das Risiko kleiner, einen Fehler 1. Art zu begehen.
  - Das Risiko, einen Fehler 1. Art zu begehen, und das Risiko, einen Fehler 2. Art zu begehen, stehen in einem Gegensatzverhältnis, d.h. eine **Verminderung eines Risikos führt zwangsläufig zu einer Erhöhung des anderen Risikos.** Sie laufen miteinander.
- **Praktische Konsequenzen**
  - Falls ein statistischer Hypothesentest durchgeführt werden soll, ist die **Fachhypothese** so zu formulieren, dass sie im statistischen Test als **Alternative** sichtbar wird. Denn wir können mit dem Hypothesentest vor allem schauen, ob die Daten gegen eine Nullhypothese sprechen und damit für die Alternative.
  - Nur leider geht das in vielen Fällen, wie z.B. bei den perlenbestückten Rechtecken der Shoshoni-Indianer nicht.

5.2 Vertrauensintervall („Intervallschätzung für Mittelwerte“)

- Mit der **Schätzung** haben wir den plausibelsten Wert für die unbekannten Koeffizienten μ und σ aus den Daten bestimmt. Dies nennt man **Punktschatzung.**
- Durch das Testverfahren haben wir gesehen, dass noch weitere Werte für μ plausibel (=mit den Daten verträglich) sein können.
- **Die Menge aller Werte für μ, die plausibel auf dem Signifikanzniveau α und somit vertraglich mit den Daten sind, bilden das so genannte (1 – α)·100%-Konfidenzintervall.**
- **Dieses Intervall lässt sich berechnen mit:** $\bar{x} \pm q_\alpha \cdot \frac{s}{\sqrt{n}}$



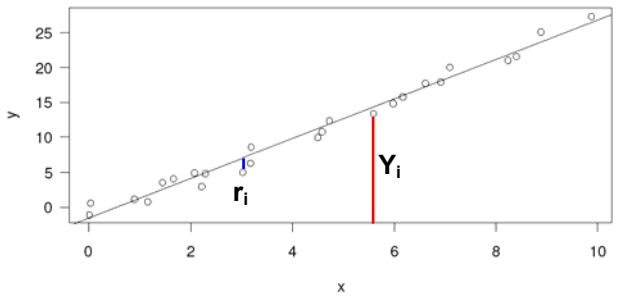
  - q_α steht für das (1 – α/2)-Quantil der t_{n-1}-Verteilung.
  - Beispiel Shoshoni-Indianer:
    - □ = 0.635, s = 0.050, n = 18
    - Signifikanzniveau α = 5 %, also beide Enden der t_{17}-Verteilung mit 2.5 % + 2.5 %. Gesucht sind also das 0.025-Quantil und das 0.975-Quantil der t_{17}-Verteilung.
      (1 – α/2) = 0.975 auflösen α = 0.05 = TINV(0.05;17) = 2.1098
    - $\bar{x} \pm q_\alpha \cdot \frac{s}{\sqrt{n}} = 0.635 \pm 2.1098 \cdot \frac{0.05}{\sqrt{18}} = [0.6101356,\ 0.6598644] \cong [0.610,\ 0.660]$
- **Gemäss der Konstruktion von Vertrauensintervallen, kann also ein Hypothesentest anstatt mit dem Einstichproben-t-Test auch mit dem Vertrauensintervall durchgeführt werden.**
  - Liegt der entsprechende Null-Hypothesen-Wert **im** (1-α) · 100 %-Vertrauensintervall, so ist die Nullhypothese **plausibel.** (d.h. die Nullhypothese kann auf dem Niveau α nicht verworfen werden.)
  - Liegt der entsprechende Null-Hypothesen-Wert **ausserhalb** des (1-α) · 100 %-Vertrauensintervalls, so ist die Nullhypothese **unplausibel.** (d.h. die Nullhypothese wird auf dem Niveau α verworfen und die Alternative gilt.)
  - Im obigen Beispiel mit den Shoshoni-Indianern liegt der hypothetische Wert 0.618 (goldener Schnitt) innerhalb des 95-%-Vertrauensintervalls, also ist die Nullhypothese auf dem 5 %-Signifikanzniveau plausibel.
- **Zwei Eigenschaften von Intervallschätzungen:**
  - **Beobachtung für n:**
    - Je grösser die Stichprobe, also die Anzahl Merkmalsträger n ist, **desto kleiner ist der Intervall.** Der Intervall läuft von beiden Seiten gegen eine bestimmte Zahl. Je mehr Daten desto genauer das Resultat.
  - **Beobachtung für α:**
    - Je grösser das Signifikanzniveau **α ist, desto kleiner ist der Vertrauensintervall.** Wählt man also ein sehr kleines Signifikanzniveau α, so ist das Vertrauensintervall sehr gross.
- **Eine weitere Interpretation der Vertrauensintervalle geht wie folgt:**
  - **Das (1-α) · 100 %-Vertrauensintervall überdeckt mit einer Wahrscheinlichkeit von (1-α) den wahren gesuchten Wert.**
  - Nehmen wir an, dass 18 betrachtete Realisierungen von N(0.618, 0.05) sind. Daraus können wir das 80 %-Vertrauensintervall bestimmen. Dann können wir nochmals 19 neue Realisierungen von N(0.618, 0.05) betrachten und wieder das 80 % Vertrauensintervall bestimmen. Dies können wir solange wiederholen bis wir 50 80 %-Vertrauensintervalle haben. **Wir erwarten dann, dass im Durchschnitt 10 Intervallen den Wert 0.618 nicht überdecken werden (20 % von 50).**



- **5.3 Was ist der Unterschied zwischen dem plausiblen Bereich und dem Vertrauensintervall?**
  - **Plausibler Bereich (Einstichproben-t-Test)**
    - Beim Einstichproben-t-Test verfügen wir über eine Nullhypothese: $Ã_0$
    - Aus den Daten bestimmen wir Ë und s.
    - Anschliessend bilden wir mit $\frac{\bar{x}-\mu}{s/\sqrt{n}}$ (**) die t_{n-1}-Verteilung.
    - Aufgrund dieser t_{n-1}-Verteilung erhalten wir einen plausiblen Bereich.
    - Es stellt sich nun die Frage ob sich $Ã_0$ innerhalb dieses plausiblen Bereichs befindet.
    - Mit dem Einstichproben-t-Test können wir nur eine Nullhypothese $Ã_0$ überprüfen.
  - **Vertrauensintervall**
    - Das Vertrauensintervall ist grundsätzlich auf dem Einstichproben-t-Test aufgebaut.
    - Wir sammeln alle Ã die plausibel sind. Das heisst, der Vertrauensintervall umfasst alle Werte für μ die auf dem Signifikanzniveau α und somit vertraglich mit den Daten sind.
    - Alle Werte μ des Vertrauensintervalls könnten wir in einem Einstichproben-t-Test nochmals überprüfen. Deren Hypothese würde immer bejaht.

6. Regressionsmodell

- **6.1 Einfache Regressionsrechnung**
  - **Methode der kleinsten Quadrate**
    - Mit der Methode der kleinsten Quadrate wird eine best-passende Gerade in ein Streudiagramm gelegt.
    - Mittelwert: $\bar{x} = \frac{\sum_{i=1}^{n} (x_i)}{n}$, $\bar{y} = \frac{\sum_{i=1}^{n} (y_i)}{n}$
    - Steigung: $b = \frac{\sum_{i=1}^{n}(y_i-\bar{y})(x_i-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$ ▪ Achsenabschnitt: $a = \bar{y} - b \cdot \bar{x}$
    - Trendgerade: $\hat{y} = a + b \cdot x_i$



  - Die Trendgerade $\hat{y}$ kann auch mit Hilfe einer Schätzung der unbekannten **Parameter β_0 (β_0 = a Achsenabschnitt)** und **β_1 (β_1 = b Steigung)** im so genannten **Regressionsmodell** betrachtet werden:
    $Y_i = \beta_0 + \beta_1 \cdot x_i + E_i$
    - Y_i wird **Zielvariable** genannt.
    - x_i ist die **erklärende Variable.**
    - E_i repräsentiert die **Abweichung von der Regressionsgeraden** und wird oft als Fehler (Error) bezeichnet.
  - Es ist nahe liegend, den Fehler E_i als eine Zufallsvariable zu betrachten. Das Kriterium der kleinsten Quadrate ist optimal, wenn E_i normalverteilt ist mit Erwartungswert 0 und Standardabweichung σ. σ ist auch unbekannt und muss geschätzt werden:
    $s = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(r_i)^2}$ mit $r_i = y_i - (a + b \cdot x_i)$
    - r_i ist der **Effektive Fehler** zwischen dem effektiven y_i und dem mit der Regressionsgerade bestimmten $\hat{y} = a + b \cdot x_i$
  - Wie beim vorhergehenden Modell strueen auch hier die Schätzungen, wenn man das Experiment wiederholen kann.
  - Ein Output eines **Statistik-Programmes wie SPSS** könnte wie folgt aussehen:

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.68172   0.39389   -4.27 0.000288
x            2.84398   0.07561   37.62  < 2e-16

Residual standard error: 1.077 on 23 degrees of freedom
Multiple R-Squared: 0.984,      Adjusted R-squared: 0.9833
F-statistic:  1415 on 1 and 23 DF,  p-value: < 2.2e-16
```

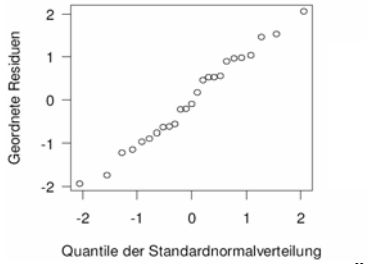
- **Was ist was?**
  - -1.68172 — **Schätzung von β_0 (β_0 = a Achsenabschnitt)**
  - 2.84398 — **Schätzung von β_1 (β_1 = b Steigung)**
  - 1.077 — **Schätzung von σ (σ = s, residual standard error, Standardabweichung)**
  - 0.39389 — **Schätzung der Streuung (Standardabweichung) von a**
    0.07561 — **Schätzung der Streuung (Standardabweichung) von b**

- **95 %-Vertrauensintervall für β_0:**

  -1.68172 ± **q** · 0.39389 = [-2.497; -0.867]

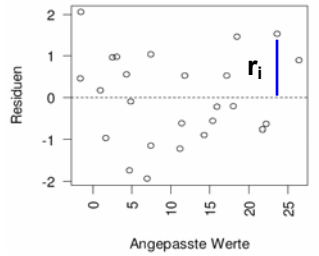
  wobei **q** das 0.975-Quantil der t_{n-2} ($=t_{23}$) Verteilung ist.
  Diese Angaben sind nur gültig, wenn die Fehler normalverteilt sind.

- **6.2 Überprüfung der Annahmen**
  - Die Grösse $r_i = y_i - (a + b \cdot x_i)$ wird **Residuum** genannt und kann verwendet werden, um die **Normalitätsannahme zu überprüfen: Quantil-Quantil-Plot von r_i;s**



Der Q-Q-Plot zeigt eine gute Übereinstimmung mit der Normalverteilung.

  - Es muss auch sicher gestellt sein, dass die Punkte **tatsächlich um eine Regressionsgerade streuen** und nicht etwa z.B. um eine Parabel. Dies kann mit einem **Streudiagramm „r_i gegen $(a + b \cdot x_i)$“** (r_i ist auf der vertikalen Achse) überprüft werden.



Die Abstände zwischen den Punkten und der Linie zeigen die Abweichungen r_i. Da die Daten **strukturlos** um die horizontale Gerade bei 0 streuen, können wir davon ausgehen, dass die Daten um eine Gerade streuen.