5. Univariate Datenbeschreibung

Nach Aufbereitung der Rohdaten liegt ein auswertbarer Datensatz, z.B. lung1984.dta, vor und man kann jetzt, zuerst eindimensional (univariat), charakteristische Kennwerte, Häufigkeiten und Verteilungen der Variablen bestimmen.
Nach dem Start des Programms wird das Stata-Dokument lung1984.dta geöffnet und eine kurze Beschreibung der Daten mit ***describe*** angefordert.

```
. use "D:\Stata\lung1984.dta"

. describe

Contains data from D:\Stata\lung1984.dta
  obs:           102
 vars:             5                          6 Jun 2021 10:46
 size:         2,040
-------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
-------------------------------------------------------------------------------
num             float   %9.0g                 Student number
sex             float   %9.0g                 Sex (female=1, male=2)
ht              float   %9.0g                 Height (cm)
pefr            float   %9.0g                 PEFR (litre/min)
vc              float   %9.0g                 Vital capacity (litre)
-------------------------------------------------------------------------------
```

Die Beispieldaten enthalten 102 Beobachtungen und 5 numerische Variable. Die Datei in dieser Form wurde am 6.Juni 2021 erstellt und belegt einen Speicher von 2040 bytes. Der voreingestellte Speichertyp für Zahlen zwischen $-1{,}7\cdot10^{38}$ und $1{,}7\cdot10^{38}$ ist "float", bei dem jede Zahl 4 bytes belegt. Bei 102 Beobachtungen und 5 Variablen ergibt sich ein Speicherplatz von $(4+4+4+4+4)\cdot102 = 2040$ bytes. Möchte man nur von einigen Var Informationen anfordern, gibt man z.B. ***describe ht pefr*** ein. Weitere Informationen zu den Var erhält man auch z.B. mit ***codebook sex ht***

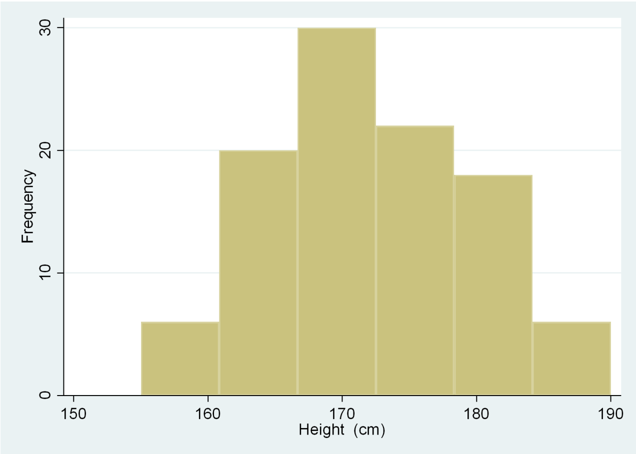
```
sex                                                     Sex (female=1, male=2)
       type:  numeric (float)
       range:  [1,2]                        units:  1
       unique values:  2                    missing .:  0/102
       tabulation:  Freq.  Value
                     44      1
                     58      2
------------------------------------------------------------------------------
ht                                                                Height  (cm)
       type:  numeric (float)
       range:  [155,190]                    units:  .1
       unique values:  46                   missing .:  0/102
       mean:   172.344
       std. dev:   8.35545
       percentiles:     10%       25%       50%       75%       90%
                        161     166.6    171.75       178     183.6
```

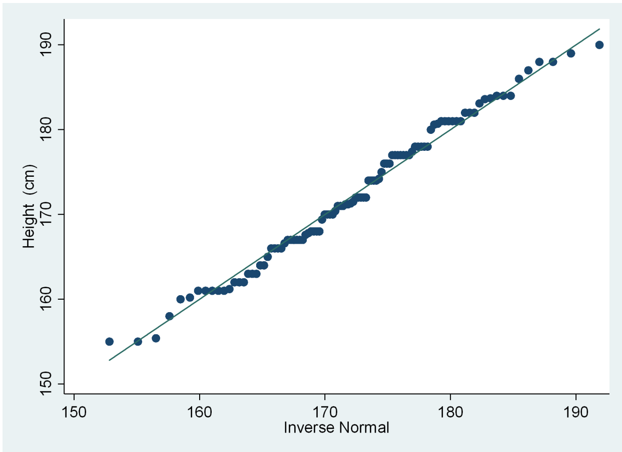
Mit ***inspect ht*** erhält man eine angedeutete Verteilung von ***ht*** sowie die Anzahl der ganzzahligen und nicht ganzzahligen Werte.
Die Kommandos ***describe, codebook und inspect*** lassen sich auch im Menü unter *Data >> Describe Data* aufrufen.

```
. inspect ht

ht:  Height  (cm)                                  Number of Observations
________________                            _________________________________________
                                                     Total     Integers   Nonintegers
 |              #           Negative                    -            -             -
 |        #  #  #           Zero                        -            -             -
 |        #  #  #           Positive                  102           83            19
 |        #  #  #                              __________   __________    __________
 |     #  #  #  #  #        Total                     102           83            19
 |     #  #  #  #  #        Missing                     -
 +----------------------                       __________
155               190                                 102
  (46 unique values)
```

5.1. Häufigkeiten

Zur Darstellung der Häufigkeitsverteilung z.B. der Variablen ***ht*** eignet sich ein Histogramm, das man mit ***histogram ht, frequency*** erhält. Es gibt, wie in (4) beschrieben, eine Reihe weiterer Möglichkeiten zur Bearbeitung von Grafiken.

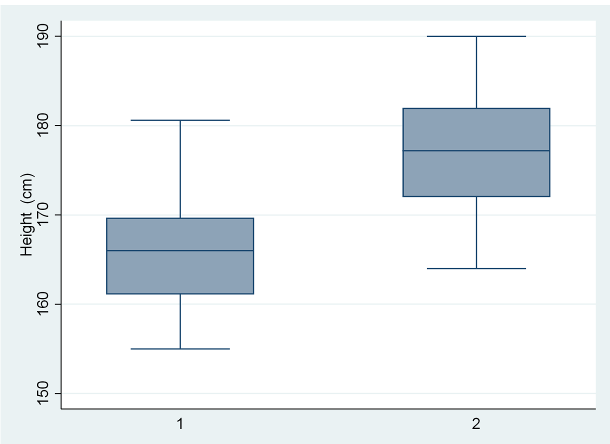




Mit ***qnorm ht*** erhält man einen Q-Q-Plot zur grafischen Beurteilung einer vermuteten Normalverteilung von **ht**.

graph box ht, over (sex)

liefert die Box Plots für **ht** seperat für Frauen (1) und Männer (2).



Eine tabellarische Darstellung der Häufigkeitsverteilung nominaler oder ordinaler Var erhält man z.B. mit ***tabulate***. Mit diesem Kommando lassen sich eindimensionale und zweidimensionale Häufigkeitstabellen erstellen. Als Kurzform für ***tabulate: tab***

Zur Demonstration rufen wir geschl_karies.dta auf. Mit der Langform erhält man:
tabulate geschl die eindimensionale Häufigkeitstabelle für ***geschl***

```
tabulate geschl
```

geschl	Freq.	Percent	Cum.
1	41	51.90	51.90
2	38	48.10	100.00
Total	79	100.00	

und mit dem Wertelabel ***label define Geschlecht 1 "W" 2 "M"*** (siehe Abschnitt 3.)

```
tabulate geschl
```

geschl	Freq.	Percent	Cum.
W	41	51.90	51.90
M	38	48.10	100.00
Total	79	100.00	

Im Vorgriff auf Abschnitt 6. erhält man eine zweidimensionale Häufigkeitstabelle (Vierfeldertafel, Kreuztabelle) für **geschl** und **karies**

tabulate geschl karies

```
tabulate geschl karies
```

	karies		
geschl	nein	ja	Total
W	14	27	41
M	12	26	38
Total	26	53	79

5.2. Erläuterungen zu den Kurzformen von *tabulate*

Für das Kommando ***tabulate*** gibt es zwei Kurzformen: ***tab1*** und ***tab2***

Mit ***tab1*** erhält man nur 1-dim. Tabellen, für jede Var eine Tabelle:

```
tab1 geschl
```

```
-> tabulation of geschl
```

geschl	Freq.	Percent	Cum.
1	41	51.90	51.90
2	38	48.10	100.00
Total	79	100.00	

oder mit zwei Variablen:

```
tab1 geschl karies
```

```
. tabulation of geschl
```

geschl	Freq.	Percent	Cum.
1	41	51.90	51.90
2	38	48.10	100.00
Total	79	100.00	

```
tabulation of karies
```

karies	Freq.	Percent	Cum.
0	26	32.91	32.91
1	53	67.09	100.00
Total	79	100.00	

Mit ***tab2*** gibt es nur 2-dim. Tabellen (Vierfeldertafeln) und daher für ***tab2 geschl*** eine Fehlermeldung:

```
. tab2 geschl
too few variables specified
r(102);
```

Für ***tab2 geschl karies*** erhält man die obige Vierfeldertafel (2x2 - Tafel)

```
tab2 geschl karies
```

```
. tabulation of geschl by karies
```

geschl	karies 0	1	Total
1	14	27	41
2	12	26	38
Total	26	53	79

Nimmt man eine dritte Var dazu in der Art ***tab2 geschl karies pla*** erhält man insgesamt 3 Vierfeldertafeln ..(geschl - karies) , (geschl - pla) , (karies - pla). Würde man 4 Var hinzunehmen ***tab2 v1 v2 v3 v4 ,*** so erhält man 12 Vierfeldertafeln ....usw.

Zusammenfassung der *tab* - Regeln

Seien v_1 v_2 v_3 ... v_k Variablen in einem Datensatz, dann erhält man mit:

	v_1	v_1 v_2	v_1 v_2 v_3 ... v_k
tab	1 HT	1 KT	Fehlermeldung
tab1	1 HT	2 HT	k HT
tab2	Fehlermeldung	1 KT	$\binom{k}{2}$ KT

Bezeichnungen: HT - Häufigkeitstabelle, KT - Kreuztabelle, $\binom{k}{2}$ Binomialkoeffizient

Zur Darstellung der Häufigkeit von metrischen Daten sollte zuerst eine Gruppierung durchgeführt werden (siehe Punkt 4).

5.3. Lage- und Streuungsmaße

Zur Demonstration öffnen wir lung1984.dta in Stata.

Häufig gebrauchte Maßzahlen metrischer Var berechnet man z.B mit ***summarize***
summarize (var) (if) (in) (weight) (, options)

Beispiel:
summarize ht Abkürzung für summarize: ***sum***

```
. sum ht
```

Variable	Obs	Mean	Std. Dev.	Min	Max
ht	102	172.3441	8.355452	155	190

Für die Variable ***ht*** werden die Zahl der Beobachtungen, Mittelwert, Standardabweichung, Minimum und Maximum berechnet. Ergebnisse für Subgruppen, wie z.B. sex, erhält man mit dem Präfix ***bysort***
bysort sex: sum ht was gleichbedeutend ist mit ***by sex, sort: sum ht***

```
. by sex, sort: sum ht
```

```
-> sex = 1
```

Variable	Obs	Mean	Std. Dev.	Min	Max
ht	44	165.8068	5.860328	155	180.6

```
-> sex = 2
```

Variable	Obs	Mean	Std. Dev.	Min	Max
ht	58	177.3034	6.307194	164	190

Eleganter in einer Tabelle geht es mit: ***tab sex, sum (ht)***

```
tab sex, sum (ht)
```

Sex (female=1, male=2)	Summary of Height (cm) Mean	Std. Dev.	Freq.
1	165.80682	5.8603283	44
2	177.30345	6.3071938	58
Total	172.34412	8.3554523	102

Eine detailliertere Berechnung für ***ht*** liefert
sum ht, detail oder für Subgruppen ***bysort sex: sum ht , detail***

```
. sum ht, detail
```

Height (cm)

Percentiles		Smallest		
1%	155	155		
5%	160.2	155		
10%	161	155.4	Obs	102
25%	166.6	158	Sum of Wgt.	102
50%	171.75		Mean	172.3441
		Largest	Std. Dev.	8.355452
75%	178	188		
90%	183.6	188	Variance	69.81358
95%	186	189	Skewness	.0309148
99%	189	190	Kurtosis	2.242026

Eine kompakte Darstellung vieler Maßzahlen einer metrischen Var liefert z.B. für ***ht***
tabstat ht, stats(n, mean, sum, sd, var, cv, sem, sk, kur, max, min, range) by(sex)

```
. tabstat ht, stats(n, mean, sum, sd, var, cv, sem, sk, kur, max, min, range) by(sex)

Summary for variables: ht
     by categories of: sex (Sex (female=1, male=2))
```

sex	N	mean	sum	sd	variance	cv	se(mean)	skewness	kurtosis	max	min	range
1	44	165.8068	7295.5	5.860328	34.34345	.0353443	.8834777	.2980303	2.79732	180.6	155	25.60001
2	58	177.3034	10283.6	6.307194	39.78069	.0355729	.8281751	-.0206508	2.233127	190	164	26
Total	102	172.3441	17579.1	8.355452	69.81358	.0484812	.8273131	.0309148	2.242026	190	155	35

Konfidenzintervalle für Mittelwerte erhält man entweder über
mean ht

```
. mean ht

Mean estimation                 Number of obs    =       102
```

	Mean	Std. Err.	[95% Conf.	Interval]
ht	172.3441	.8273131	170.703	173.9853

oder mit

ci means ht

```
. ci means ht
```

Variable	Obs	Mean	Std. Err.	[95% Conf.	Interval]
ht	102	172.3441	.8273131	170.703	173.9853

Die Berechnung von Konfidenzintervallen für Anteile erhält sie man über:

***proportion* sex**

```
. proportion sex, citype(normal)

Proportion estimation                 Number of obs    =       102
```

	Proportion	Std. Err.	Normal [95% Conf.	Interval]
sex				
1	.4313725	.0490388	.3340927	.5286524
2	.5686275	.0490388	.4713476	.6659073

Dabei gibt es verschiedene Möglichkeiten der Schätzung von Konfidenzintervallen bei denen geringe Abweichungen in der Breite dieser Intervalle auftreten können:

Confidence interval type

- ○ Logit
- ○ Exact (Clopper-Pearson)
- ◉ Normal (Wald)
- ○ Wilson
- ○ Agresti-Coull
- ○ Jeffreys

Handelt es sich beim Studiendesign um Clusterstichproben, so führt dies in der Regel zu einer Varianzvergrößerung und damit zu breiteren Konfidenzintervallen. Die Datei **kiga_5.dta** ist das Ergebnis einer Zufallsstichprobe von 5 Kindergärten, in denen alle Kinder zahnärztlich untersucht wurden (einstufige Clusterstichprobe). Der dmf - Wert zeigt die Zahl der kariesbefallenen Zähne pro Kind. Ohne Berücksichtigung des Clusterdesigns erhält man mit ***mean dmf*** für den mittleren dmf-Wert 1,93 mit einem Konfidenzintervall C.I. = (1,46 ; 2,39)

```
. mean dmf

Mean estimation                       Number of obs    =       201
```

	Mean	Std. Err.	[95% Conf.	Interval]
dmf	1.925373	.2365738	1.458874	2.391872

Wird das Clusterdesign berücksichtigt, ***mean dmf, vce(cluster kiga)***, ergibt sich ein C.I. = (1,40 ; 2,45) und somit ein breiteres Konfidenzintervall.

Nähere Ausführungen über Clusterstichproben findet man in der Rubrik "Statistik im ÖGD"

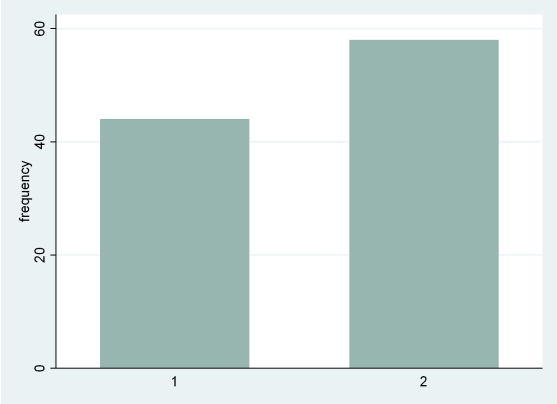
```
. mean dmf, vce(cluster kiga)

Mean estimation                   Number of obs    =       201

                    (Std. Err. adjusted for 5 clusters in kiga)
------------------------------------------------------------------------------
             |                Robust
             |       Mean   Std. Err.      [95% Conf. Interval]
-------------+----------------------------------------------------------------
         dmf |   1.925373   .1892828       1.39984    2.450906
------------------------------------------------------------------------------
```

5.4. Verteilungen

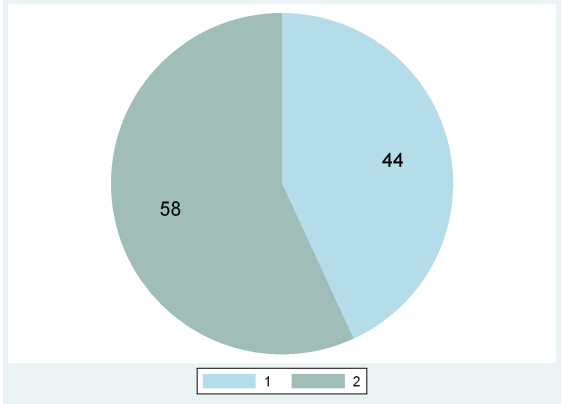
Mit ***histogram ht, bin(6) start(155) frequency***
erhält man beim Datensatz lung1984.dta die oben in 5.1. dargestellte Häufigkeitsverteilung für die Var ***ht*** in sechs Klassen. Verschiedene Boxplots dieser Var sind unter 4. Grafiken zu finden.

Für die Var ***sex*** wären Balken- oder Kreisdiagramm sinnvoll.

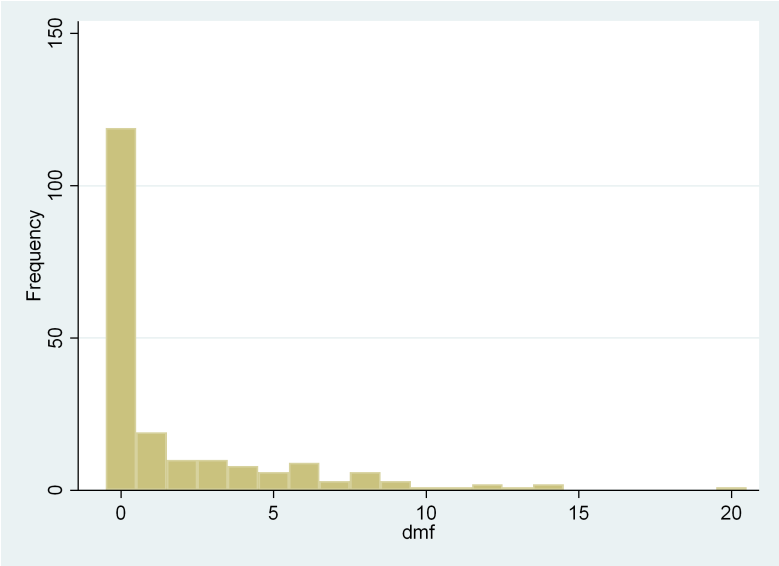
graph bar (count), over(sex)



graph pie, over(sex) plabel(_all sum)



Für die Var ***dmf*** im Datensatz kiga_5.dta findet man folgende Verteilung:
histogram dmf, discrete frequency



STATA - Kommandos für univariate Datenbeschreibung

describe

codebook

inspect

qnorm

tabulate

tab1 , ***tab2***

summarize

bysort v: sum

tabstat v1, stats(n, mean, sum, sd, var, cv, sem, sk, kur, max, min, range) by(v2)

mean

ci means

proportion