2. Descriptive Statistics

Readings - Rosner: 2.1-2.6; 2.9-2.10
SAS: PROCs UNIVARIATE, MEANS, CORR
R: writing functions; Hmisc: quantile, boxplot, qqnorm, cor

A) Measures of Center (Location)
B) Measures of Spread (Variation)
C) Percentiles, quantiles
D) General Properties of Expectations and Variances

Descriptive Statistic: a single number that summarizes one aspect of a sample so that we can make concise quantitative statements about the data. It is often convenient to summarize data in terms of a pair of numbers, one telling us where the distribution is centered and the other telling us how spread out the distribution is.

Example: As an example of the calculations, we'll study daily intake (kcal) for 13 women and 32 men from the previous dataset. Let's describe the diets of the subjects in the diet study. About how many calories did they eat per day? About how much fat did they eat? How variable were those from person to person? If a person eats 3000 calories in a day, how do they compare to this group? etc…

Notation:

Data values are denoted by: $X_1, X_2, \cdots, X_n$

The sum of those values can be denoted using summation notation: $X_1 + X_2 + \cdots + X_n = \sum_{i=1}^{n} X_i$

A) Measures of Center (Location) – Locating the “middle” of the sample

1) Arithmetic Mean (Mean): Average of the data. Sum of the data divided by the number of data points $X_1 + X_2 + \cdots + X_n$:

$$\overline{X} = \frac{X_1 + X_2 + \cdots X_n}{n} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Interpretation: Center of gravity of the distribution of the data – the point on which it would perfectly balance from left to right.

- Most widely used measure of central tendency.
- Sensitive to extreme values, as in skewed (asymmetric) distributions
- Used for continuous and discrete variables, but not usually appropriate for ordinal variables.

e.g. the mean intake for females: 836, 1196, 1340, 1352, 1588, 1708, 1760, 1855, 1902, 2212, 2313, 2821, 3086

$\bar{X}$ = (836 + 1196 + … + 3086) /13 = 1844 kcal

```
PROC SORT DATA=diet; BY sex;

ODS PDF;

      PROC MEANS DATA=diet;
       VAR kcal3 ;
       BY sex;
      RUN;

ODS PDF CLOSE;
```

The MEANS Procedure
sex=Female

Analysis Variable : kcal3				
N	**Mean**	**Std Dev**	**Minimum**	**Maximum**
13	1843.77	637.6452977	836.0000000	3086.00

sex=Male

Analysis Variable : kcal3				
N	**Mean**	**Std Dev**	**Minimum**	**Maximum**
32	2714.84	852.9350885	1435.00	4475.00

The sample mean $\bar{X}$ estimates or approximates the *population mean* μ.

$\bar{X}$ is an unbiased estimate of μ. $E\left[\bar{X}\right]=\mu$

On repeated (random) sampling from a population, with samples of size *n*, the average value of $\bar{X}$ over all possible samples is μ.

In general, an estimator is unbiased if its expected value equals the parameter of interest.

What happens to $\bar{X}$ when we translate, or add a constant *c* to the observations X_i?

$X_1, X_2, X_3, \ldots, X_n$ $\qquad$ $X_1+c, X_2+c, X_3+c, \ldots, X_n+c$

$\bar{X}$ $\qquad$ **?**

What happens to $\bar{X}$ when we rescale the observations, i.e. when we multiply them by a constant *c*?

$X_1, X_2, X_3, \ldots, X_n$ $cX_1, cX_2, cX_3, \ldots, cX_n$

$\bar{X}$ **?**

2) Median: middle value (or average of middle two values) when data are sorted

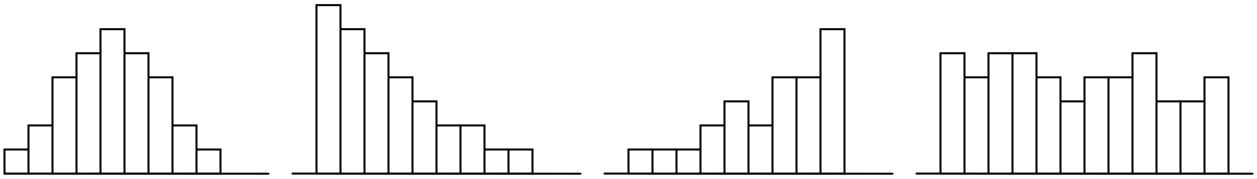
If *n* is odd the median is $\frac{n+1}{2}th$ sorted observation. If *n* is even, the median is the average of $\frac{n}{2}th$ and $\frac{n}{2}+1th$ sorted observations. Thus, there are equal numbers of observations on either side of the sample median.

e.g. the sorted calorie intake values for the females are:
836, 1196, 1340, 1352, 1588, 1708, 1760, 1855, 1902, 2212, 2313, 2821, 3086

The median food intake for females is: kcal

For some thoughts on interpreting the median, see: “The median is not the message” - **http://www.edwardtufte.com/tufte/gould**.

What is Skewness? … It's complicated … Does not exactly determine relationship of the mean to the median



Symmetric Skewed Positive Skewed Negative Uniform

In symmetric unimodal distributions, $\bar{X}$ = median = mode

e.g. height

But … skewness = 0 does not guarantee symmetry or that $\bar{X}$ = median, only that a short, fat "tail" balances out a long, thin "tail"

For multimodal distributions … all bets are off.

Unimodal, positive, right skewed, tail to the right

e.g. blood pressure, weight, cholesterol, price of housing, wages

Unimodal negative, left skewed, tail to the left

e.g. gestational age, Apgar scores, age at diagnosis for childhood cancers

For a nice summary: http://en.wikipedia.org/wiki/Skewness

3) Trimmed mean: The k% trimmed mean is the mean of the central k% of the values

e.g. the 50% trimmed mean of female food intake is:
The mean omitting 836, 1196, 1340, 2213, 2812, 3086

The 100% trimmed mean is: kcal
The 0% trimmed mean is: kcal

Note: Computer programs may use slightly different algorithms for trimming in exact percentages.

4) **Geometric mean:** antilog of the arithmetic mean of the observations that have been transformed to be on a logarithmic scale, e.g. log_2, log_{10}, or *ln*

Compute mean on the *log* scale: $\overline{\log X} = \frac{1}{n}\sum_{i=1}^{n} \log X_i$

Take the antilogarithm to convert back to original scale:
antilog of $\overline{\log X}$

The geometric mean will be the same regardless of which base is used to compute the logarithm as long as the logs and antilogs in the definition are in the same base.

Log-transformations tend to "normalize" skewed data and are good to use when observation values differ by orders of magnitude.

4) **Harmonic mean:** Typically, the harmonic mean is appropriate for situations when the average of *rates* is desired. It is used when making multiple comparisons with unequal *n*'s.

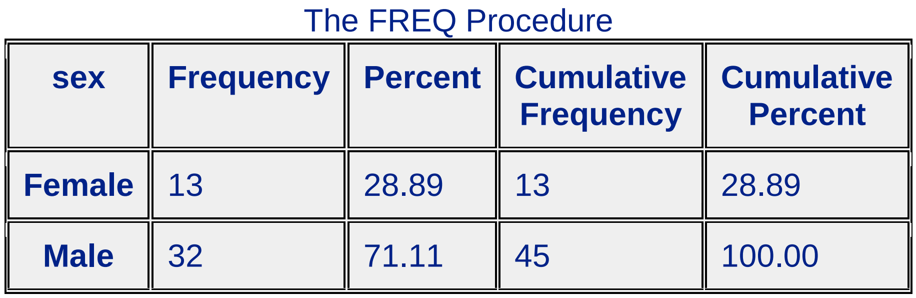
$$h=\frac{n}{\sum_{i=1}^{n}\frac{1}{X_i}} \qquad let\,Y=\sum_{i=1}^{n}\frac{1}{X_i} \qquad \bar{Y}=\frac{\frac{1}{X_1}+\frac{1}{X_2}+\ldots+\frac{1}{X_n}}{n}$$

$$h=\frac{1}{\bar{Y}}$$

5) Mode: Most common or frequently occurring value in a sample.
For symmetric distributions, mean = median = mode
In right skewed distributions, the mode is slightly left of the median. In left skewed distributions, the mode is slightly right of the median.
Biomodality may be evidence of a mixture of populations in our sample

These measures of central tendency are appropriate for continuous and discrete variables.

For nominal and ordinal variables, it's more meaningful to report frequencies and percentages of totals.



The FREQ Procedure

sex	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Female	13	28.89	13	28.89
Male	32	71.11	45	100.00

B) Measures of Spread (Variation or Dispersion):
How much variability is there about the point of central tendency (where the data aggregate)?

1) Range: Difference between largest observation and smallest
Largest value – smallest value: $X_{(n)} – X_{(1)}$

e.g. Calorie intake for females: 836, 1196, 1340, 1352, 1588, 1708, 1760, 1855, 1902, 2212, 2313, 2821, 3086

Range of food intake for females was: 3086 – 836 = 2250 kcal

Range is easy to compute and describe but it has some shortcomings:

--It is sensitive to extreme values.

--It tends to increase with increasing sample size making it difficult to compare samples with different *n*'s.

2) Interquartile Range (IQR):

Lower quartile: Q_1 = median of data below the median
Upper quartile: Q_3 = median of data above the median
(if *n* is odd, the median is usually included in both)
IQR = Upper quartile – Lower quartile = $Q_3 – Q_1$ = 75th percentile - 25th percentile

e.g. Sorted values of calorie intake for females: 836, 1196, 1340, 1352, 1588, 1708, 1760, 1855, 1902, 2212, 2313, 2821, 3086

For female food intake: Q_1 = Q_3 =
IQR = kcal = kcal

3) Variance:

An intuitive estimator of variability about the mean is the average of the *deviations* of the observations from the mean. Unfortunately the average will always be 0:

$$\sum_{i=1}^{n}\left(X_i-\bar{X}\right)=\sum_{i=1}^{n}X_i-\sum_{i=1}^{n}\bar{X}=n\bar{X}-n\bar{X}=0$$

We might then be tempted to take the *absolute values* of the deviations and then average those. That would work, but the resulting quantity is not as mathematically tractable as the following.

We can define the sample variance, $\hat{\sigma}^2$, as the average of the *squared* deviations of the observations from their mean:

$$\hat{\sigma}^2=\frac{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2}{n}$$ (∧ = “hat”, and denotes “estimate of”)

This seemingly fine estimator also has one weakness: it’s *not* an *unbiased* estimator of *population* variance σ^2**.** It is, however, the maximum likelihood estimate, which you will learn about next semester if you’re enrolled in BIOS 6632.

$$E\left[\hat{\sigma}^2=\frac{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2}{n}\right]\neq\sigma^2$$

For example, if we were to take all possible random samples of size 25 from a population, the average value of $\hat{\sigma}^2$ would not be equal to σ^2.

In order to have an *unbiased estimate* of σ^2 we need to divide by *n-1* instead of by *n*:

$$s^2 = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}{n-1} \quad , \quad E\left[s^2 = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}{n-1} \right] = \sigma^2$$

n-1 can also be thought of as the *degrees of freedom* – we lose one degree of freedom when we estimate the population mean by $\bar{X}$.

The larger *n* is, the less difference the denominator makes since *n* is closer to *n-1*.

Another formula for the sample variance that can be useful for hand calculations and sometimes in proofs: $s^2 = \dfrac{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2}{n-1}$

Proof: $(n-1)s^2 = \sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 = \sum_{i=1}^{n} X_i^2 - n\bar{X}^2 = \sum_{i=1}^{n} X_i^2 - \dfrac{\left(\sum_{i=1}^{n} X_i\right)^2}{n}$

This is known as the *computational formula*. It can be less stable numerically but only takes one pass through the data whereas the original formula takes two.

Note: some calculators divide by n, not n-1

4) Standard Deviation:

Population standard deviation (s.d.) = $\sigma = \sqrt{\sigma^2}$

Sample standard deviation (s.d.) = $s = \sqrt{s^2} = \sqrt{\dfrac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$

s is not an unbiased estimator of the population s.d. σ**:** $E[s] \neq \sigma$ - we use it anyway!

e.g. for female food intake, $\bar{X}$ = 1844 kcal

i	X_i	$X_i - \bar{X}$	$(X_i - \bar{X})^2$
1	1855	11	121
2	1340	-504	254016
⋮	⋮	⋮	⋮
13	1352	-492	242064

$$\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 = 4879092$$

variance = s^2 = 4879092/12 = 406591 kcal^2

s.d. = s = 637.65 kcal

The more variable the data are, the more spread out the distribution is and the greater the SD and variance are.

The variance is easier to work with mathematically, but the standard deviation is easier to interpret in practice, mainly because units of s^2 are (data units)^2 and units of s are (data units) Neither of these quantities is meaningful when dealing with skewed distributions.

For symmetric distributions that appear "normal" (more on this in a couple of weeks), the s.d. can be visualized as follows:

About 68% of the area under the curve is within *one* s.d. of the mean.

About 95% of the area under the curve is within *two* s.d. of the mean.

The *normal ranges* often used for diagnostic purposes in clinical medicine are generally calculated as $\bar{X} \pm 2s$.

If the data have a normal or Gaussian distribution $\bar{X} \pm 2s$ will correspond to the 2.5 and 97.5 percentiles and the normal range will include about 95% of the normal group.

(If the data are highly skewed, Chebychev's inequality guarantees that $\bar{X} \pm 2s$ will include at least 75% of the distribution.)

What happens to s^2 and s when we translate the data?

$X_1, X_2, X_3, \ldots, X_n$	$X_1+c, X_2+c, X_3+c, \ldots, X_n+c$
s	?
s^2	?

Both are unaffected by translation:

$$s^2=\frac{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2}{n-1} \qquad s_c^2=\frac{\sum_{i=1}^{n}\left(X_i+c-\left(\bar{X}+c\right)\right)^2}{n-1}$$

What happens to s^2 and s when the data are rescaled?

$X_1, X_2, X_3, \ldots, X_n$	$cX_1, cX_2, cX_3, \ldots, cX_n$
s	?
s^2	?

Variability is affected by the amount rescaled:

$$s_c^2=\frac{\sum_{i=1}^{n}\left(cX_i-c\bar{X}\right)^2}{n-1}=s_c^2=\frac{\sum_{i=1}^{n}c^2\left(X_i-\bar{X}\right)^2}{n-1}=c^2s^2$$

$$s_c=\sqrt{s_c^2}=\sqrt{c^2s^2}=cs$$

Different *sources* of variation can be quantified by different standard deviations. Note that a simple standard deviation makes sense only when there is one source of variability.

e.g. Three replicate determinations of insulin clearance rate on each of 4 dogs:

Dog	1st Det	2nd Det	3rd Det	Mean Dog
1	75	65	72	70.7
2	96	92	76	88.0

3	98	109	99	102.0
4	91	97	99	95.7
Mean Det	90	90.75	86.5	

For any given dog, a s.d. can be calculated based on the 3 replications made on that dog, which measures the variability between replications (perhaps attributable to laboratory technical error in the assay procedures). This is called *within-subject* variability.

For any determination (or for the mean determinations) a s.d. can be calculated based on 4 measurements which measures the variability from dog to dog, perhaps attributable to differences between dogs. This is called *between-subject* variability

Basing a s.d. on all 12 measures would be mixing two sources of variability.

5) Coefficient of Variation (CV): $CV = \left(\frac{s}{\bar{X}}\right) or \left(\frac{s}{\bar{X}}\right)$x100%

Useful for comparing variation of different variables (with different means and/or units of measurement) or for comparing variation of two samples for the same variable. Also useful in situations where variability tends to increase with the magnitude of the observations. Usually CV is used only for positive variables.

e.g. which is more variable in weight, cows or mice?

e.g. Pop 1 avg DBP is much larger than avg DBP in Pop 2 e.g. in single population, is cholesterol or DBP more variable—use CV to determine which is more variable relative to its mean.

e.g. kcal intake

	s	$\bar{X}$	CV
Male	853	2715	.314
Female	638	1844	.346

C) **Percentiles (or Quantiles):** *p* percent of a distribution is less than the *p*th percentile (quantile); calculation depends on value of $\frac{np}{100}$

--If $\frac{np}{100} \neq$ integer: *p*th percentile is the *(k+1)*th largest X_i where *k* is the largest integer < $\frac{np}{100}$

$X_1, X_2, X_3, \ldots, X_n$, n = 50

$X_1', X_2', X_3', \ldots, X_{50}'$ (ordered)

95th percentile of this distribution: $\frac{np}{100} = \frac{50(.95)}{100} = 47.5 =$

k = 47, k+1 = 48 X_{48}' = 95th percentile

--If $\frac{np}{100}$ = integer: pth percentile is the average of the $\frac{np}{100}th$ and $\left(\frac{np}{100}+1\right)th$ largest observation

90th percentile of this distribution: $\frac{np}{100} = \frac{50(.90)}{100} = 45$

$\frac{X'_{45} + X'_{46}}{2} = 90^{th}$ percentile

Spread of distribution can be characterized by specifying several percentiles (e.g. 10th and 90th). In medical applications, percentiles (usually 2.5 and 97.5 percentiles) are often used as so-called "normal limits".

Percentiles are less sensitive to outliers and not as greatly affected by sample size. Note: algorithms and results for computing percentiles may vary among software

Median =	50th percentile
Q1 =	25th percentile
Q3 =	75th percentile

Q. A variable has the percentiles: 10th = 4, 25th = 9, 50th = 15, 75th = 30, 90th = 85.

Describe the distribution …

D) Correlation: Used to describe the relationship between two continuous variables (e.g., height and weight). It measures the *strength* (qualitatively) and *direction* of the linear relationship between two or more variables.

- The Pearson correlation coefficient measures the strength of the ***linear*** association between two variables, X and Y. It can be used to estimate the population correlation, ρ, and is defined by:

$$r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2 \sum_{i=1}^{n}(Y_i - \overline{Y})^2}}$$

- A correlation can be between –1 and 1:
  - If the correlation is greater than 0, then as *X* increases *Y* increases and the two variables are said to be ***positively correlated***. An *r* = 1 is perfect positive correlation.
  - If the correlation is less than 0, then as *X* increases *Y* decreases and the two variables are said to be ***negatively correlated***. An *r* = -1 is perfect negative correlation.
  - If the correlation is 0 then there is no linear relationship between *X* and *Y*. The two variables are said to be ***uncorrelated***.

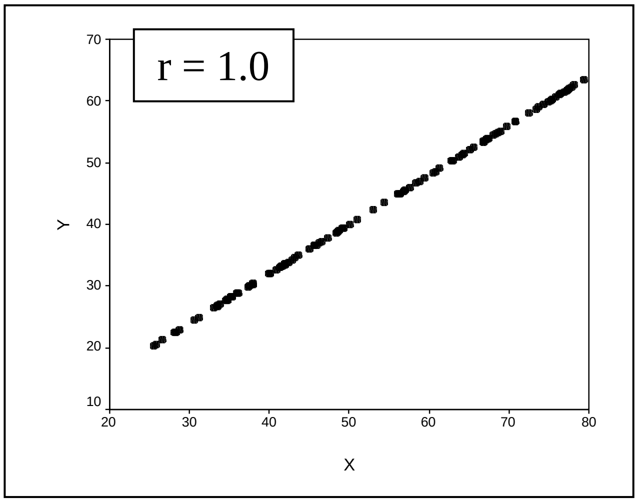
- Spearman's rank correlation coefficient is a rank-based analogue to Pearson's correlation coefficient. It can be used when the assumption of (bivariate) normality is not met – more on normality later …

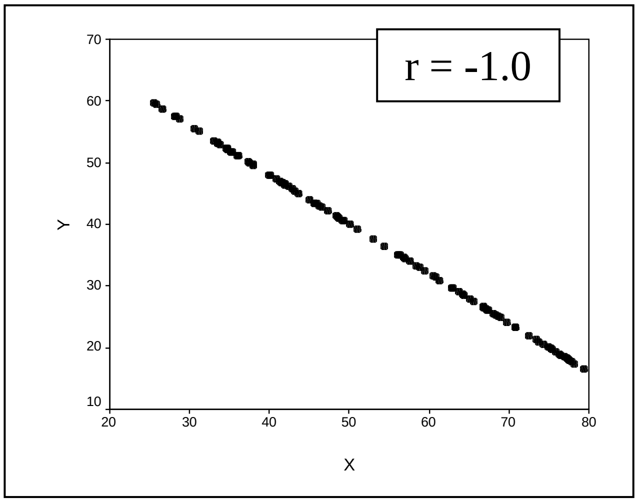
- Spearman's rank correlation is based on the ranked data.
  (1) Rank the values for each measurement separately.
  (2) Compute Pearson's correlation coefficient on the ranked data.

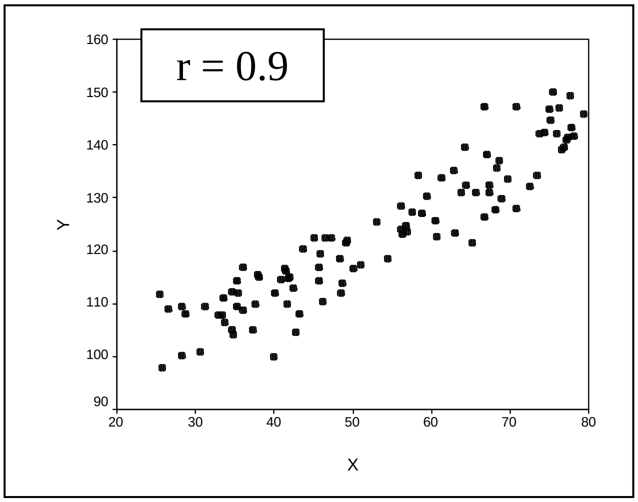
$$-1 \leq r_S \leq 1$$

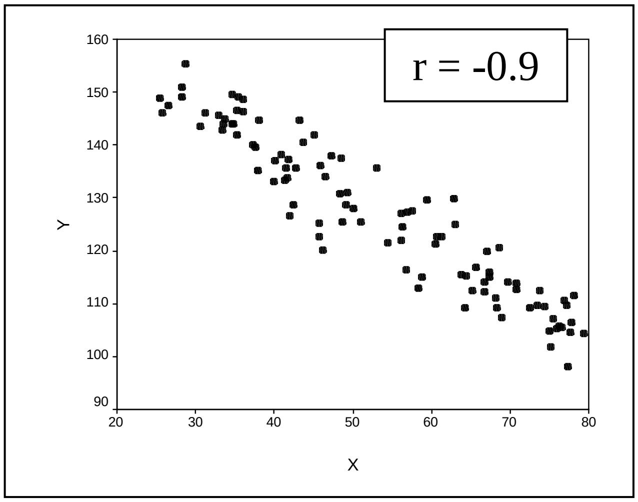
- When r and r_s differ a lot, there is an indication that the variables are not normally distributed; r_s should then be used.

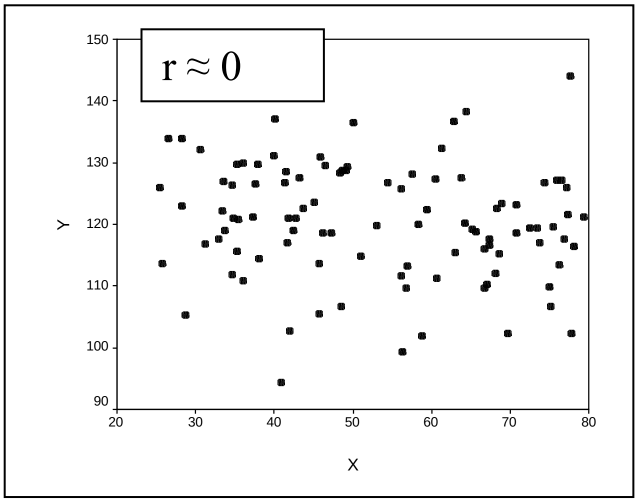
- Spearman’s rank correlation is a measure of monotonic association (not necessarily linear).


r = 1.0
Y
70
60
50
40
30
20
10
20
30
40
50
60
70
80
X


r = -1.0
Y
70
60
50
40
30
20
10
20
30
40
50
60
70
80
X


r = 0.9
Y
160
150
140
130
120
110
100
90
20
30
40
50
60
70
80
X


r = -0.9
Y
160
150
140
130
120
110
100
90
20
30
40
50
60
70
80
X


r ≈ 0
Y
150
140
130
120
110
100
90
20
30
40
50
60
70
80
X

Uses and Misuses of Correlation

from Altman, D.G. *Practical Statistics for Medical Research*, Chapman and Hall, 1991

- Spurious correlations involving time
  - When both variables have time trends, this will induce a (potentially spurious) correlation between them.

- Restricted sampling of individuals
  - Between-subject variability makes a direct impact on the calculation of the correlation coefficient.

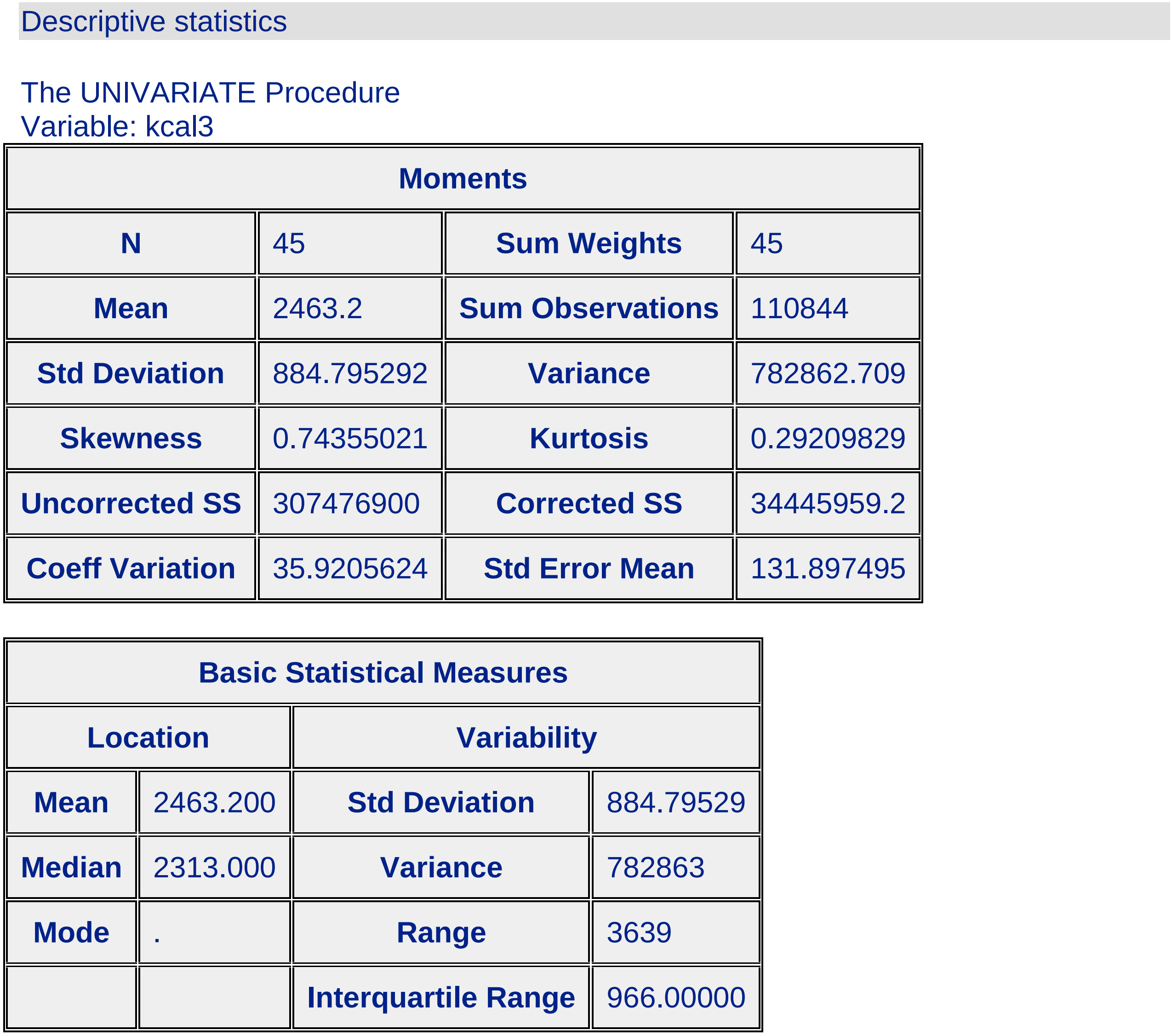
- Mixed samples
  - The presence of subgroups with different associations may result in an overall correlation that is not representative of the correlation within the subgroups.

- Assessing agreement
  - Correlation measures association, not how closely two measurements agree (i.e. how much variability there is between two or more measures on the same unit).

- Change related to initial value
  - Regression to the mean (recall the research replication study)

- Relating a part to a whole
  - X and X + Y will always appear to be correlated.

```
ODS PDF;
    PROC UNIVARIATE DATA=diet PLOT;
     VAR kcal3;
     TITLE 'Descriptive statistics';
    RUN;

    PROC CORR DATA=diet PEARSON SPEARMAN;
     VAR fdwt3 kcal3;
     TITLE 'Food weight vs. Calorie intake';
    RUN;
ODS PDF CLOSE;
```



Descriptive statistics

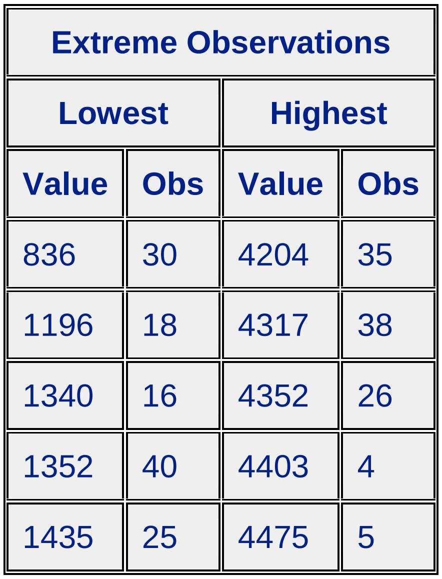
The UNIVARIATE Procedure
Variable: kcal3

Moments			
N	45	**Sum Weights**	45
Mean	2463.2	**Sum Observations**	110844
Std Deviation	884.795292	**Variance**	782862.709
Skewness	0.74355021	**Kurtosis**	0.29209829
Uncorrected SS	307476900	**Corrected SS**	34445959.2
Coeff Variation	35.9205624	**Std Error Mean**	131.897495

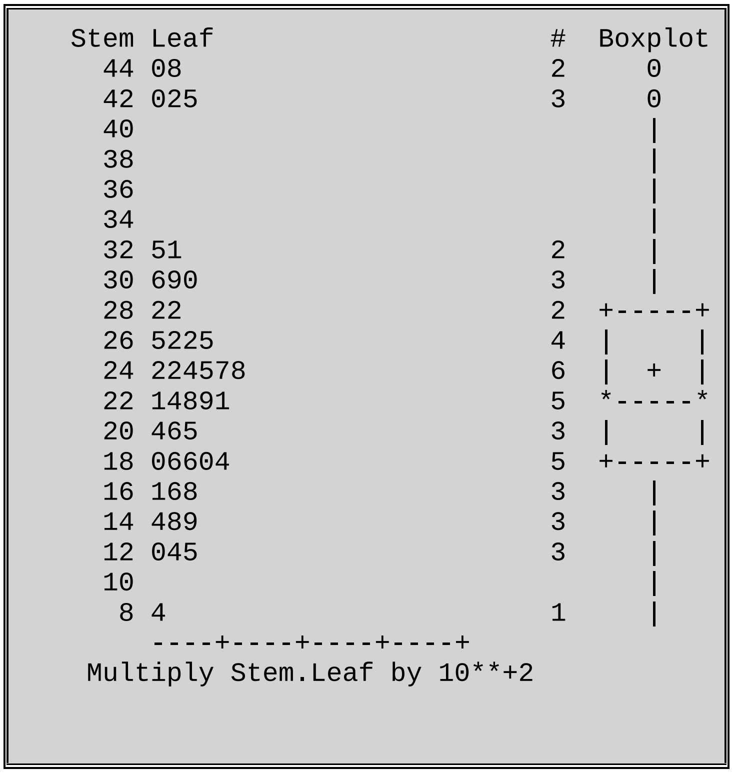
Basic Statistical Measures			
Location		**Variability**	
Mean	2463.200	**Std Deviation**	884.79529
Median	2313.000	**Variance**	782863
Mode	.	**Range**	3639
		Interquartile Range	966.00000

Tests for Location: Mu0=0				
Test	**Statistic**		**p Value**	
Student's t	**t**	18.67511	**Pr > |t|**	<.0001
Sign	**M**	22.5	**Pr >= |M|**	<.0001
Signed Rank	**S**	517.5	**Pr >= |S|**	<.0001

Quantiles (Definition 5)	
Quantile	**Estimate**
100% Max	4475
99%	4475
95%	4352
90%	4204
75% Q3	2821
50% Median	2313
25% Q1	1855
10%	1435
5%	1340
1%	836
0% Min	836



Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
836	30	4204	35
1196	18	4317	38
1340	16	4352	26
1352	40	4403	4
1435	25	4475	5



```
 Stem Leaf                     #  Boxplot
   44 08                       2     0
   42 025                      3     0
   40                                |
   38                                |
   36                                |
   34                                |
   32 51                       2     |
   30 690                      3     |
   28 22                       2  +-----+
   26 5225                     4  |     |
   24 224578                   6  |  +  |
   22 14891                    5  *-----*
   20 465                      3  |     |
   18 06604                    5  +-----+
   16 168                      3     |
   14 489                      3     |
   12 045                      3     |
   10                                |
    8 4                        1     |
      ----+----+----+----+
  Multiply Stem.Leaf by 10**+2
```

Descriptive statistics

The UNIVARIATE Procedure
Variable: kcal3

```
                       Normal Probability Plot
    4500+                                           *   *++
        |                                      ** *   ++
        |                                           ++
        |                                         ++
        |                                      +++
        |                                    ++
        |                                  ++**
        |                                +***
        |                             +++*
    2700+                           ++***
        |                         +***
        |                       ***
        |                    +**
        |                  ***
        |               ****
        |            ***+
        |         * *+
        |       * ++
     900+   *   ++
         +----+----+----+----+----+----+----+----+----+----+
             -2        -1         0        +1        +2
```

Pearson Correlation Coefficients, N = 45 Prob > |r| under H0: Rho=0		
	fdwt3	**kcal3**
Fdwt3	1.00000	0.83670 <.0001
Kcal3	0.83670 <.0001	1.00000

Spearman Correlation Coefficients, N = 45
Prob > |r| under H0: Rho=0

	fdwt3	**kcal3**
fdwt3	1.00000	0.83584 <.0001
kcal3	0.83584 <.0001	1.00000

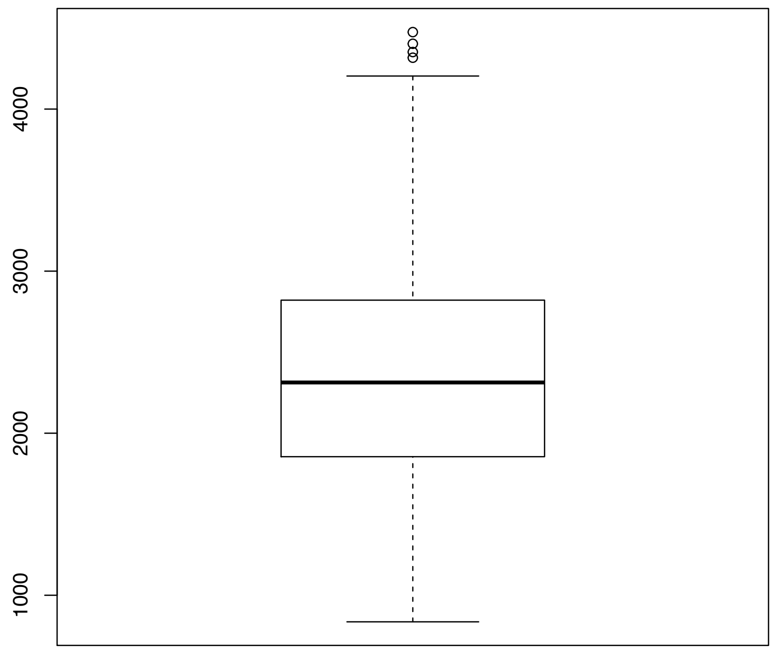
☺

Did you hear about the statistician who had his head in an oven and his feet in a bucket of ice? When asked how he felt, he replied,

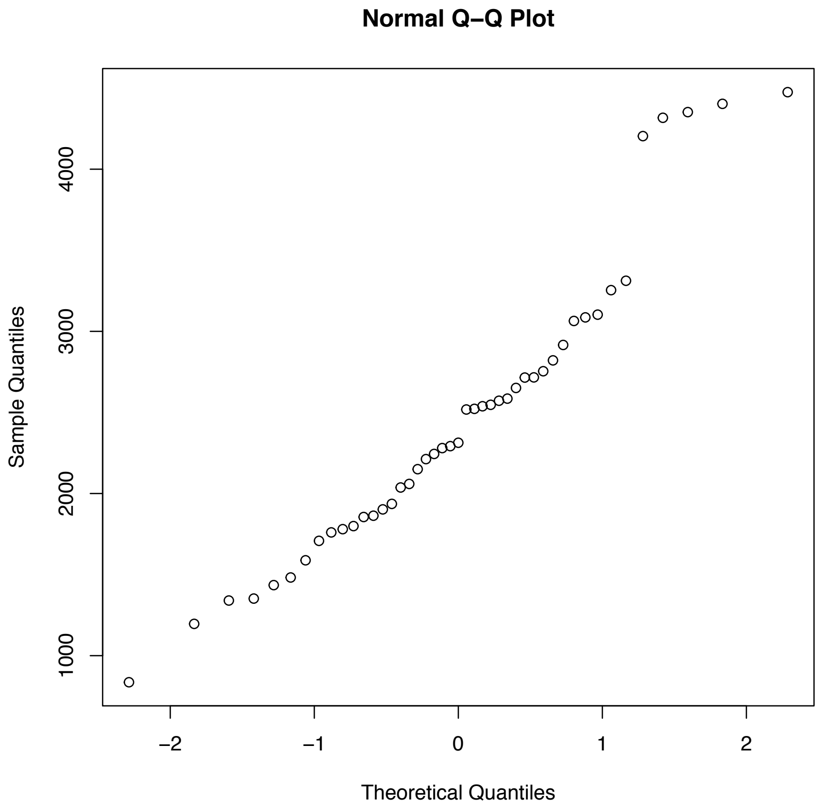
"On the average I feel just fine."

```
diet <- read.csv("~/Dropbox/6611METHODS/6611/diet.csv")
diet$gender <- factor(diet$sex, levels = c(0, 1), labels = c("female", "male"))
x.fun <- function(x) {
    x.mean <- mean(x)
    x.sd <- sd(x)
    x.min <- min(x)
    x.max <- max(x)
    out <- c(x.mean, x.sd, x.min, x.max)
    names(out) <- c("mean", "sd", "min", "max")
    return(out)
}
with(diet, tapply(kcal3, gender, x.fun))

$female
  mean     sd    min    max 
1843.8  637.6  836.0 3086.0 

$male
  mean     sd    min    max 
2714.8  852.9 1435.0 4475.0 

k <- diet$kcal3
summary(k)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
    836    1860    2310    2460    2820    4480 

library(Hmisc)
describe(k)

k 
      n missing  unique    Mean     .05     .10     .25     .50     .75 
     45       0      45    2463    1342    1454    1855    2313    2821 
    .90     .95 
   3847    4345 

lowest :  836 1196 1340 1352 1435, highest: 4204 4317 4352 4403 4475

sd(k)

[1] 884.8

var(k)

[1] 782863

range(k)

[1]  836 4475
```

```
t.test(k)


One Sample t-test

data:  k
t = 18.68, df = 44, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 2197 2729
sample estimates:
mean of x
     2463

wilcox.test(k)


Wilcoxon signed rank test

data:  k
V = 1035, p-value = 5.684e-14
alternative hypothesis: true location is not equal to 0

quantile(k)

  0%  25%  50%  75% 100%
 836 1855 2313 2821 4475

quantile(k, probs = c(0, 0.01, 0.05, 0.1, 0.25, 0.5, 0.75, 0.9, 0.95, 0.99,
    1))

    0%     1%     5%    10%    25%    50%    75%    90%    95%    99%
 836.0  994.4 1342.4 1453.8 1855.0 2313.0 2821.0 3847.2 4345.0 4443.3
  100%
4475.0

sk <- sort(k)
head(sk)

[1]  836 1196 1340 1352 1435 1482

tail(sk)

[1] 3312 4204 4317 4352 4403 4475

boxplot(k)
```



```
stem(k)


  The decimal point is 3 digit(s) to the right of the |

  0 | 8
  1 | 2344
  1 | 5678889999
  2 | 01222333
  2 | 555566777889
  3 | 11133
  3 |
  4 | 2344
  4 | 5

qqnorm(k)
```



```
f <- diet$fdwt3
cor(k, f)

[1] 0.8367

cor.test(k, f)


Pearson's product-moment correlation

data:  k and f
t = 10.02, df = 43, p-value = 8.167e-13
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.7200 0.9074
sample estimates:
   cor
0.8367

cor(k, f, method = "spearman")
```

```
[1] 0.8358

cor.test(k, f, method = "spearman")


Spearman's rank correlation rho

data:  k and f
S = 2492, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
   rho
0.8358
```

Appendix: Code

```
diet <- read.csv("~/Dropbox/6611METHODS/6611/diet.csv")
diet$gender<- factor(diet$sex,levels = c(0,1),labels = c("female", "male"))
x.fun<-function(x){
  x.mean<-mean(x)
  x.sd<-sd(x)
  x.min<-min(x)
  x.max<-max(x)
  out<-c(x.mean,x.sd,x.min,x.max)
  names(out)<-c("mean","sd","min","max")
  return(out)
}
with(diet,tapply(kcal3,gender,x.fun))
with(diet,xtabs(~kcal3+gender))

k<-diet$kcal3
summary(k)
library(Hmisc)
describe(k)
sd(k)
var(k)
range(k)
t.test(k)
wilcox.test(k)
quantile(k)
quantile(k,probs=c(0,0.01,0.05,0.1,0.25,0.50,0.75,0.9,0.95,0.99,1))
sk<-sort(k)
head(sk)
tail(sk)
boxplot(k)
stem(k)
qqnorm(k)
f<-diet$fdwt3
cor(k,f)
cor.test(k,f)
cor(k,f,method="spearman")
cor.test(k,f,method="spearman")
```