Lectures 4

Normality checks and non-parametric mean tests

Non-parametric tests

- t-tests assume that variables are normally distributed

But:

is not bell-shaped

1) sometimes variable ***distribution is not normal***

2) or ***sample is too small*** (i.e. there are too few cases in the sample to allow reliable estimation of normal parameters μ and σ)

- In such cases, *non-parametric tests* must be used instead of t-tests

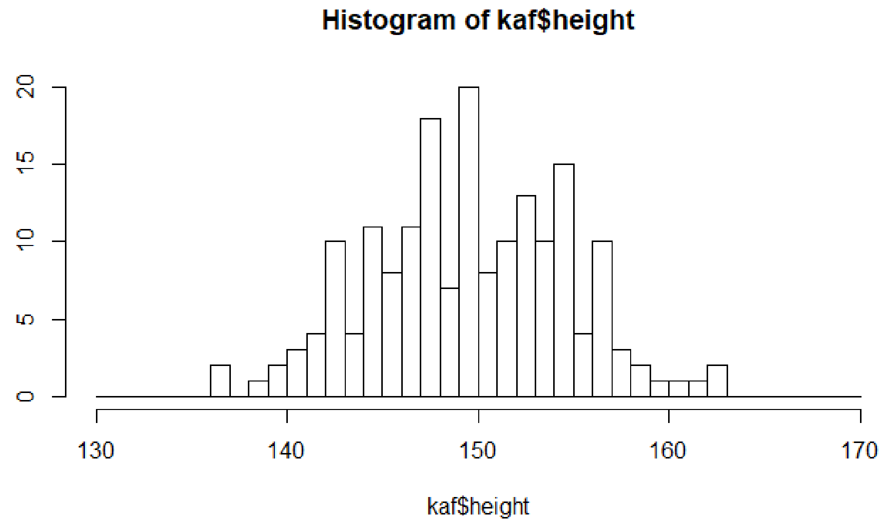
- This lecture introduces
  - normality tests
  - non-parametric alternatives to *t*-tests

Checking for normality

Relates to common sense

How can you check for normality?

- For example, take adult female height in the !Kung

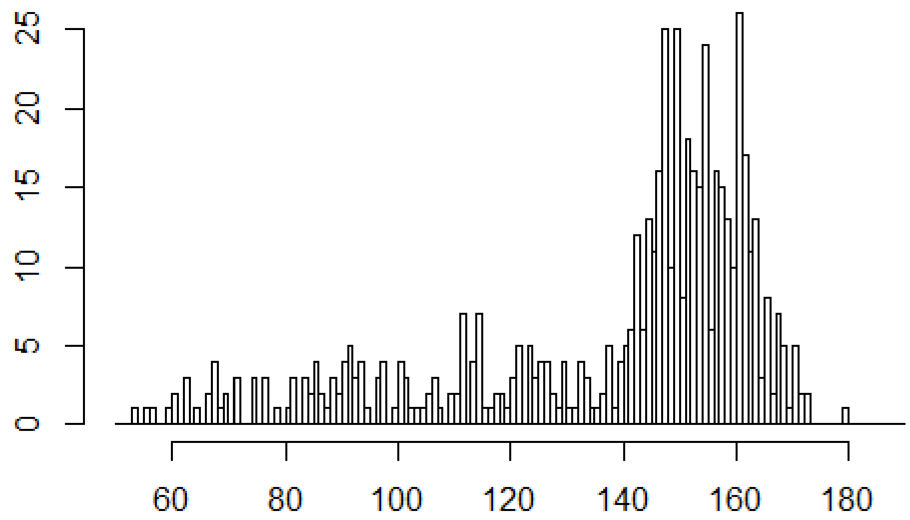


(i) *Visual inspection*

- Look for bell-shaped histogram
- Best and most direct indication of normal distribution

Checking for normality

- What about all heights of all (!Kung) women and men, children and adults)?
- Distribution of !Kung height is not normal
  - because of children, curve has a long tail below mean
  - clear indication of non-normal distribution
- Visual check should be followed by formal testing for normality



formal test

Shapiro-Wilk test

(ii) Visual check should be followed by formal normality tests

- as a rule, they compare observed sample values to values predicted from normal distribution with the same observed mean and sd

- The *Shapiro-Wilk test* calculates *W* statistics (Kendall's tau) that measures concordance between observed (sample) vs. predicted (normal curve) values

- Null hypothesis: variable is normally distributed
  - =no significant difference to a normal distribution with same mean and sd
  - If *P*>0.05, variable is normal — Accept H0
  - If *P*<0.05, variable is not normal = significant difference — Reject H0

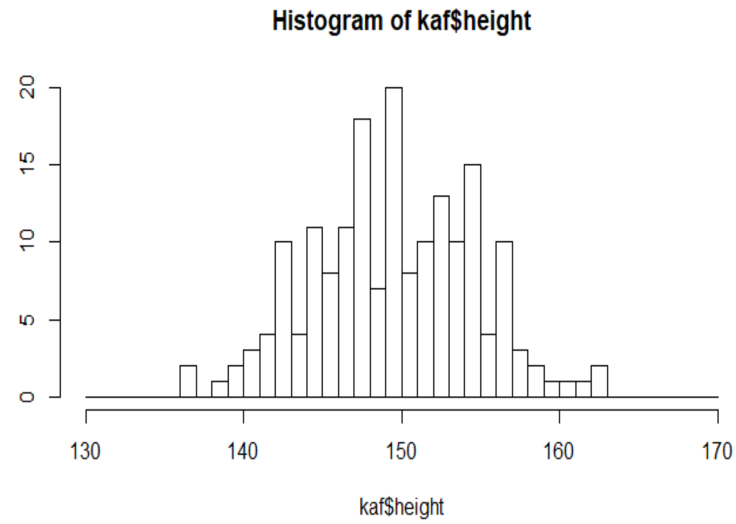
Shapiro-Wilk test

- Example: !Kung adult female heights
- Histogram looked bell-shaped; but is distribution normal?

```
> shapiro.test(kaf$height)

        Shapiro-Wilk normality test

data:  kaf$height
W = 0.99401, p-value = 0.6761
```



- P=0.68
- null hypothesis cannot be rejected at a significance level of P=0.05
- !Kung adult female height is normally distributed

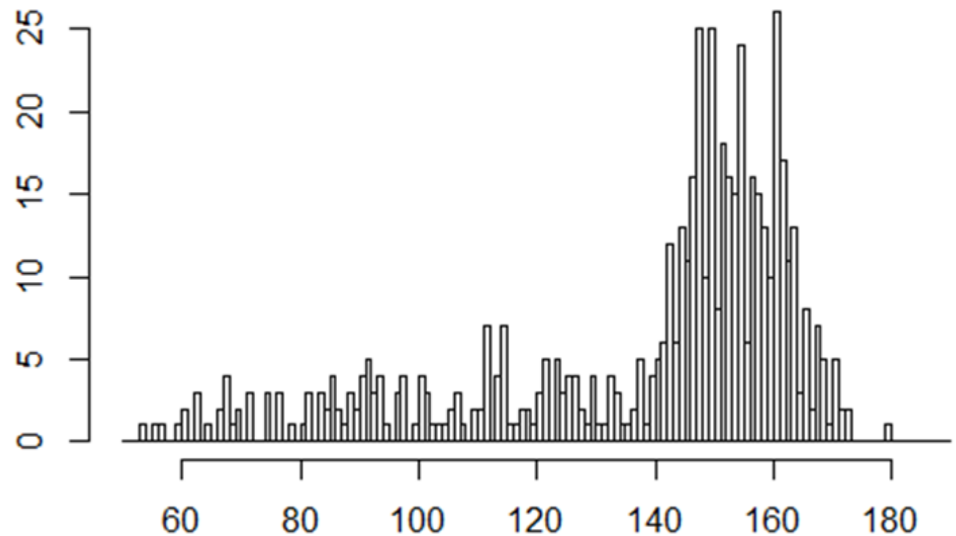
Shapiro-Wilk test

- What about height of all !Kung?

```
> shapiro.test(kc$height)

        Shapiro-Wilk normality test

data:  KungCensus$height
W = 0.8383, p-value < 2.2e-16
```



Conclusion: *reject* null hypothesis

- !Kung height (adults + children) is not normally distributed

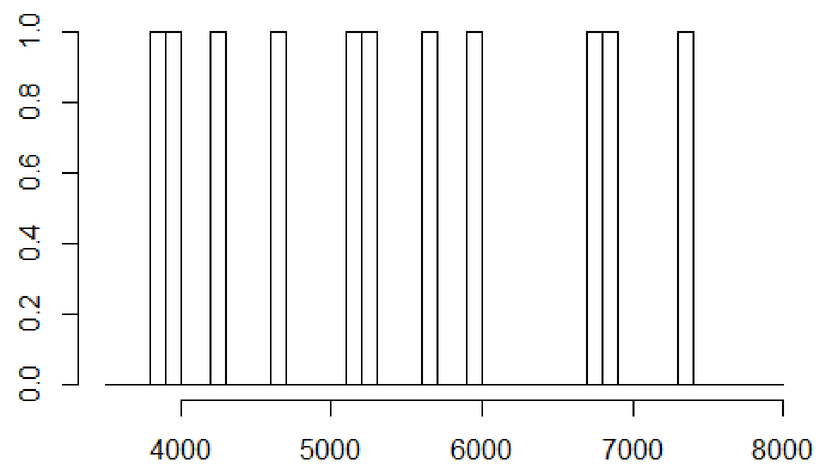
Small samples

- **Important**: previous examples are based on large samples
  - there are enough data to reject null hypothesis of normality
  - how much is ‘large enough’? No clear answer; over 20-30? 10-20 cases may be too few cases

- Post-menstrual calories intake (*intake$post*, library *ISwR*), with N=11
  - histogram does not suggest normal pattern

```
> shapiro.test(intake$post)

   Shapiro-Wilk normality test
data: intake$post
W = 0.9364, p-value = 0.4787
```

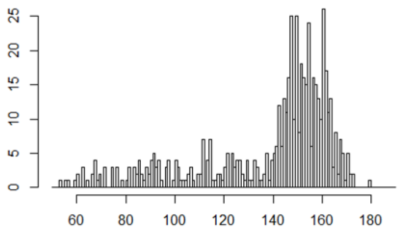


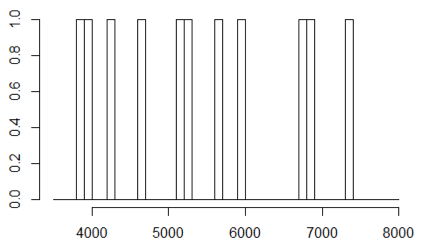
- **Shapiro-Wilk test is not very sensitive; it fails to reject null hypothesis (=normality) when samples are ‘small’**

When to run non-parametric tests

Run non-parametric tests (instead of t-tests) if

- sample is ‘large’ and fails Shapiro-Wilks normality test
- sample is too ‘small’, histograms don’t look bell-shaped, even if you cannot reject normality; it is safer!





Note: there are many other normality tests

- package *nortest* with lillie.test (Lilliefors aka Kolmogorov-Smirnov test) among others; same problem of small sample size applies to them

Exercise:

The file *react* (ISwR library) has differences in measurements made by two nurses

- Visualise the distribution of *react* using a basic histogram; does it look normal?
- Now divide the x axis into intervals of 1 unit using argument *seq*; does it look normal?
- Run a shapiro-wilks test; is distribution normal?

```
hist(react, breaks=seq(-9,8,1))
```

Non-parametric tests: ranking cases

- How to compare group means without assuming that your variable distribution is normal?

Simple idea is to **rank cases**:

- rank cases in your sample from largest to smallest (1st, 2nd, 3rd)
    - in a sample of heights, rank from tallest to shortest
- replace values with rankings
    - the tallest case becomes '1'
- then compare distribution of those ranks
    - to a test value (one-sample)
    - between groups (two-sample)
    - by individual (paired)

Wilcoxon signed-rank test

- =non-parametric alternative to one-sample *t*-test

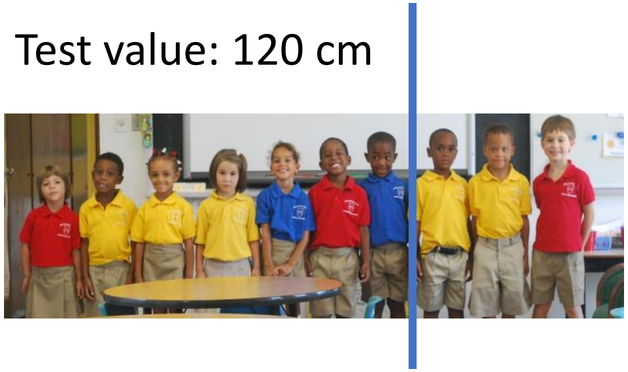
ranking and labelling from the most deviant value to the most closest value to the reference value

Example: heights in children

Do they differ in height from 120cm?

- 1. calculate and rank differences between each case and test value (disregarding sign)
  - largest difference (positive or negative) is ranked 1,
  - shortest child is 109cm tall, it is 11cm shorter than test value (120cm); shortest child receives rank=1

Test value: 120 cm



-11cm

+6cm

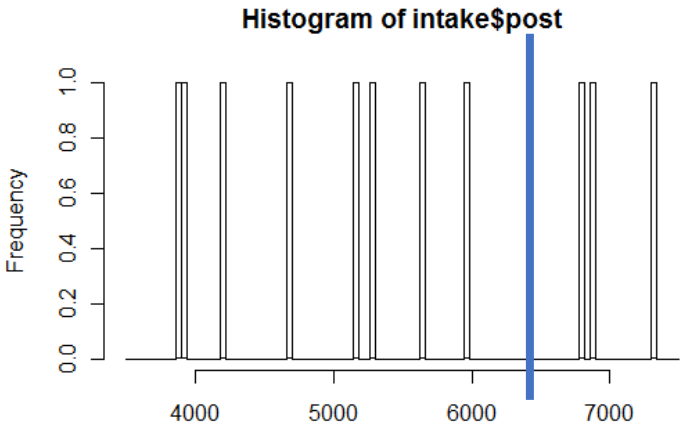
- 2. Add sign to ranks
  - If rank 1 is below test value (i.e. shorter), give it value -1; if it is taller, give it the rank +1; etc. ; shortest child receives rank=-1
- 3. Compare sum of positive vs. negative ranks
  - if sample mean is close to test value, mean of positive (taller than test value) and negative (shorter than test value) rankings should not differ much

Normally distributed values should be dispersed on two sides of test value, meaning relatively equal distribution. Otherwise, values stand in one direction.

- 4. Calculate probability of from a theoretical rank distribution (to obtain a P value)

Wilcoxon signed-rank test

- Example: is post-menstrual calorie consumption (*intake$post*) different from 6500 kcal?




```
> wilcox.test(intake$post, mu=6500, conf.int=T)
        Wilcoxon signed rank test
data:  intake$post
V = 7, p-value = 0.01855
alternative hypothesis: true location is not equal to 6500
95 percent confidence interval:
 4535 6300
sample estimates:
(pseudo)median
     5403.75
```

- V is a test statistic based on the sum of positive ranks
  - (ps. do not try to interpret V or W values; they depend in sample size and hence cannot provide a general reference for significance, such as t=±1.96 for a 95% CI)
- 'pseudomedian' is similar to median or mean
- *P*<0.05   Significant difference

Conclusion:

- =reject null hypothesis
- post-menstrual calorie consumption is significantly below 6500 kcal

Exercise:

Import file *HDR2011* (selected variables from the *Human Development Report 2011*)

- Is the distribution of the variable *HDI* (human development index) normal?
- What is the average human development index in the dataset?
- Is the average HDI in the world significantly different from 0.7? reject H0

Yes.

Two-sample Wilcoxon test

= Mann-Whitney test

- alternative to two-sample *t*-test

Similar ranking procedure:

- 1. Mix the two samples together (e.g. height in boys and girls)
- 2. Rank cases (tallest becomes 1 etc.)
- 3. Compare ranks from two samples
    - if boys and girls have similar mean heights, mean of rankings from boys and girls shouldn't differ significantly

10 9 8 7 6 5 4 3 2 1

Two-sample Wilcoxon test

- Example: do !Kung boys and girls differ in weight?

```
> wilcox.test(kb$weight ~ kb$sex, conf.int=T)
        Wilcoxon rank sum test
data:  kb$weight by kb$sex
W = 32, p-value = 0.6612
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval:
 -1.417475  3.203494
sample estimates:
difference in location
           0.56699
```

Accept H0

- W statistic is sum of ranks in first group minus minimum possible value
- difference in location = 0.57kg
- P-value: 0.66
- 95%5 CI includes 0

- = no significant difference in weight

Alternative syntax
> wilcox.test(variable1, variable2, conf.int=T)

??????????

wilcox.test(zelazo$active, zelazo$none, conf.int=T)

Exercises:

1) We use Wilcoxon tests when samples are small

- Open file *zelazo* (with data on walking age in four groups of children); read file description in the ISwR package
- Compare the groups *active* (children who received active training) and *none* (no training); which test do you use? Wilco-two sample
- Now compare *active* and control (*ctr.8w*) groups. Is there a difference?

2) Open file *energy* (with data on energy expenditure on two groups of women) from *ISwR*

- Is there a difference in energy expenditure between lean and obese women?

yes

wilcox.test(energy$expend~energy$stature,conf.int=T)

Matched-pairs Wilcoxon test

- Alternative to paired-samples *t*-test
- Example: are pre- and post-menstrual calorie consumption levels different?

```
> wilcox.test(intake$pre, intake$post, paired=T, conf.int=T)

Wilcoxon signed rank test with continuity correction
data:  intake$pre and intake$post
V = 66, p-value = 0.00384
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval:
 1037.5 1582.5
sample estimates:
(pseudo)median
     1341.332
```

V= sum of positive ranks
95% CI excludes a difference of zero
significant difference between pre- and post-consumption

Note: in Lecture 3 we applied two-sample and paired-sample t tests to this dataset
But paired-sample Wilcoxon test is the appropriate test due to small sample size!

???????

Exercise:

Look at file heart.rate (ISwR) with data on nine patients before and after taking a drug to reduce heart rates

- Is there a difference between heart rates before drug administration (time=0) and 120 days (time=120) after taking the drug?