Lectures 4

Normality checks and non-parametric mean tests

正态性检验和非参数均值检验

样本太小或者不符合正态分布不可以用t-test

Non-parametric tests

- t-tests assume that variables are normally distributed

But:

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

2) or ***sample is too small***: sample is too small to allow reliable estimation of normal parameters μ and σ

Note: to use a t-test, bith 1) and 2) must be true for ALL samples SEPARATELY

- for example, for a two-sample t-test, distributions and sample sizes of BOTH samples should be normally distributed and large
- if the two conditions are not satisfied for one of the groups, t-test is not appropriate
  - 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 检验正态性

如何检查正态分布

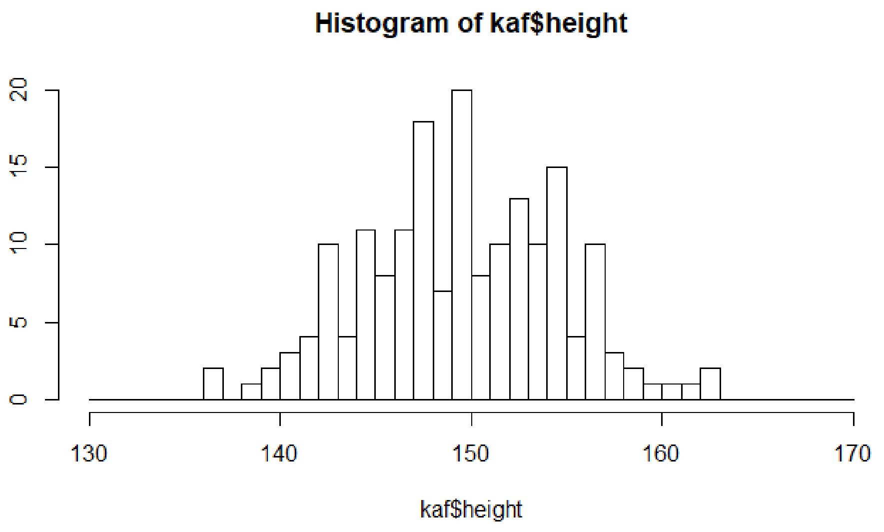
How do you check for normality?

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

(i) *Visual inspection* 目测

- Look for bell-shaped histogram
  - visual inspection is the most direct indication of normal distribution

观察钟型直方图



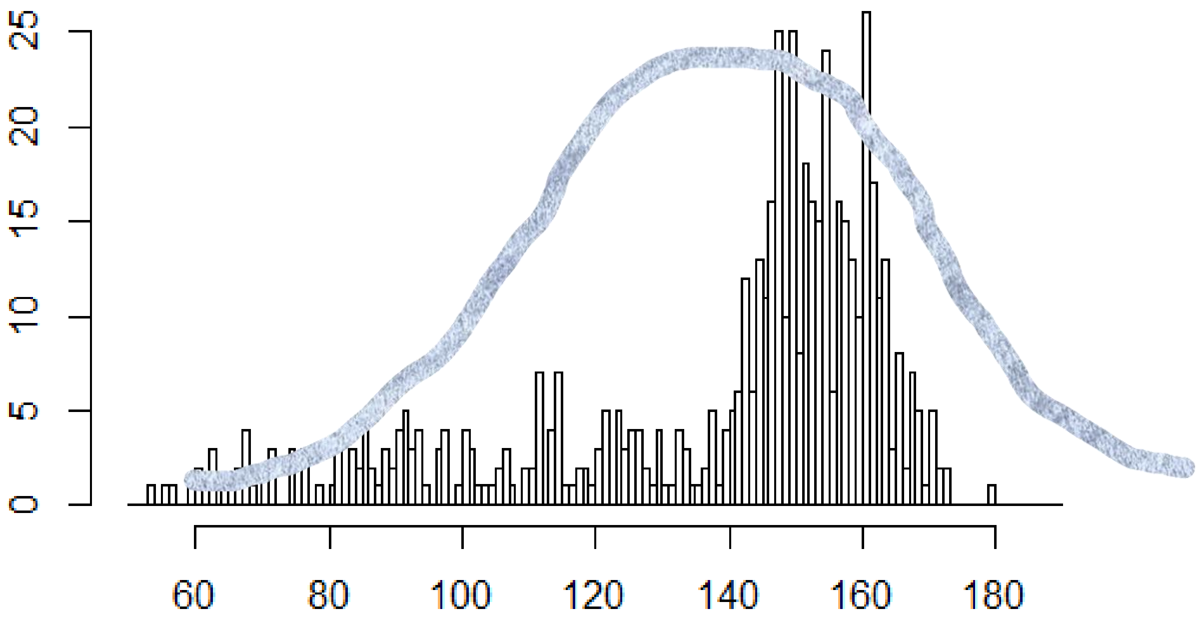
Checking for normality

- What about all heights of all (!Kung) women and men, children and adults)?

由于儿童的存在，在均值以下有很长的一侧

- Distribution of !Kung height is not normal
  - because of children, curve has a long tail below the mean
  - clear indication of non-normal distribution
  - A one-sample t-test is not appropriate in this case

在这种情况下，单样本检测的t-test是不合适的



w检测

Shapiro-Wilk test

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

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

测试将观察到的样本值与正态分布的预测值进行比较，这个预测值具有与观察到的相同的均值和方差

- The *Shapiro-Wilk test* calculates *W* statistics (Kendall's tau) that measures the match between observed (sample) vs. predicted (normal curve with same mean and sd) values

零假设：变量为正态分布=与其具有相同均值和标准差的正态分布无显著差异

- Null hypothesis: variable is normally distributed

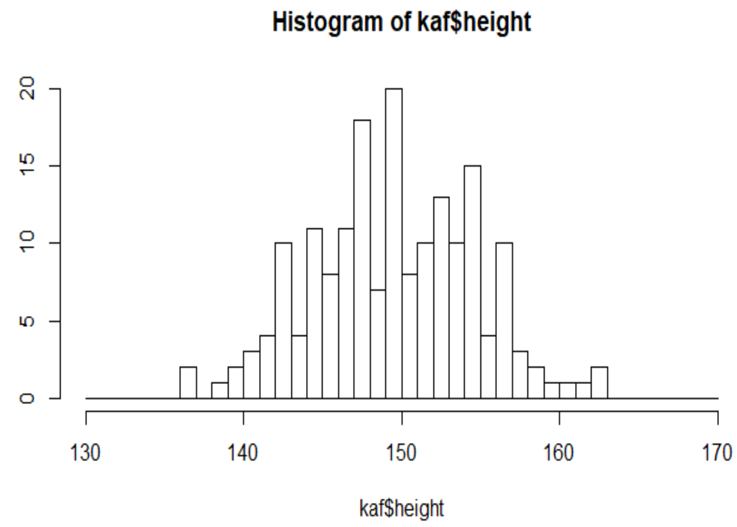
  =no significant difference to a normal distribution with same mean and sd

  - If $P>0.05$, variable is normal p值大于0.05，变量正常
  - If $P<0.05$, variable is not normal = significant difference

  小于0.05，变量不正常，有显著差异

Shapiro-Wilk test

- Example: !Kung adult female heights
- Histogram looks bell-shaped; sample size is good (n = 181);
- But is distribution normal?




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

        Shapiro-Wilk normality test

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

不知道W值代表什么意思，不看它

P=0.68 与正常曲线无显著差异

- no significant difference from normal curve
- null hypothesis cannot be rejected at a significance level of α=0.05 不能拒绝零假设，即
- = !Kung adult female height is normally distributed

成年女性身高呈现正态分布

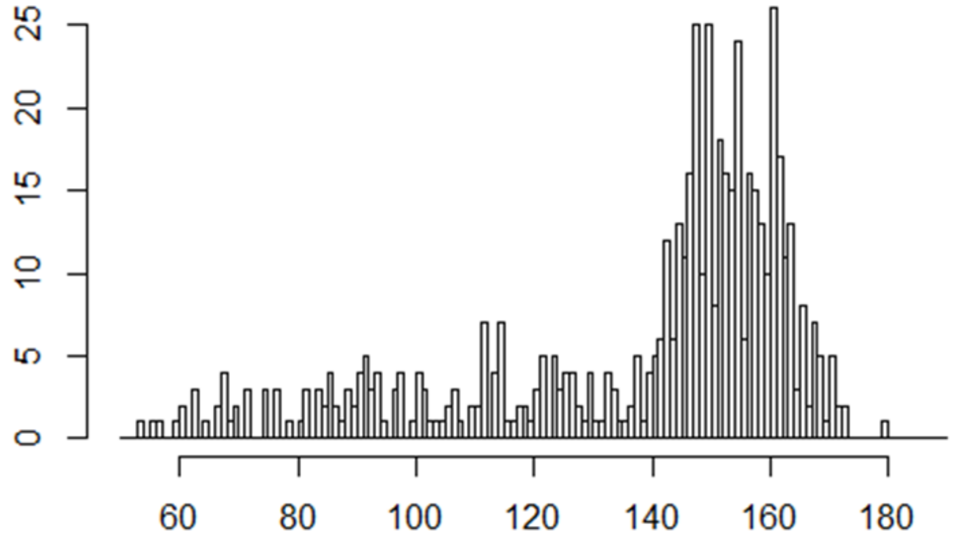
Shapiro-Wilk test

- What about height of full !Kung sample?

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

        Shapiro-Wilk normality test

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



p值远小于0.05，结论：拒绝零假设
身高(成人+儿童)是非正态分布的

P <<<<< 0.05

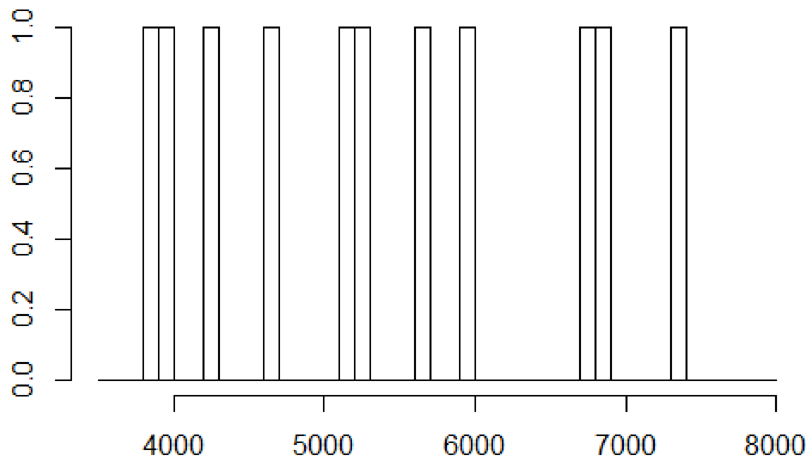
Conclusion: *reject* null hypothesis

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

Small samples 小样本

- **Important**: previous examples are based on large samples
  前面的示例是基于大样本的
  我们有足够的数据点来拒绝正态性的零假设
  - we had enough data points to reject null hypothesis of normality
  - how much is 'enough data'? No clear answer; over 20-30? 10-20 cases may be too few cases
  - For the purposes of the tests in this module, n >= 30 wil be considered large 对于本模块，大于等于30视为较大样本

- Post-menstrual calories intake (*intake$post*, library *ISwR*), with N=11
  - histogram does not suggest normal pattern
  - but test says we should consider distribution to be normal!




```
> shapiro.test(intake$post)
   Shapiro-Wilk normality test
data: intake$post
W = 0.9364, p-value = 0.4787
```

这个测试不是很敏感，当样本过小时，它不能拒绝零假设(假设样本是正态分布的)

But it is not! Shapiro-Wilk test is not very sensitive; it fails to reject null hypothesis (=normality) when samples are 'small'

何时进行非参数检验

When to run non-parametric tests

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

- sample is ‘large’ (at least 30) but Shapiro-Wilks normality test rejects normality

样本大于30但w检验出来不是正态性的

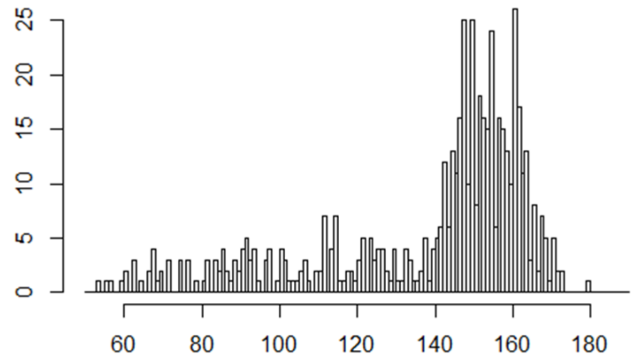
- sample is too ‘small’ (less than 30), and histograms don’t look bell-shaped, even if you cannot reject normality with Shapiro-Wilk test; it is safer!

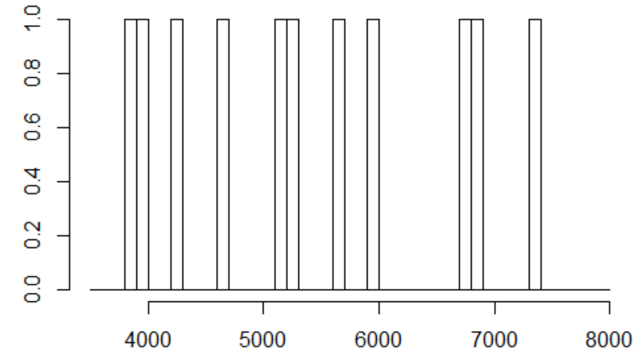
样本数量小于30，直方图看起来不是钟型的，用w检测不能排除正态性

- Again: this must apply for the two samples in the case of two-sample t-test of paired t-test

Note: there are many other normality tests 对于两个样本的也适用

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





```
> hist(react,breaks = seq(-10,10,1))
> shapiro.test(react)

 Shapiro-Wilk normality test

data:  react
W = 0.95701, p-value = 2.512e-08
```

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?

太多值在中间了，不符合正态分布

非参数检验：排序

Non-parametric tests: ranking cases

- How to compare sample means when your variable distribution is not normal?

当变量不符合正态分布，如何比较样本均值？

Simple idea is to **rank cases:** 从最大到最小排序，将值替换成排名，然后比较等级分布

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

Wilcoxon signed-rank test

- **=non-parametric alternative to the one-sample *t*-test** 单样本t检验的非参数代替方法

Example: you want to test whether height of children is significantly different from a test value (120 cm), but your sample is very small (n=10)

=> sample is too small: you cannot run one-sample t-test

计算每个孩子和测试值之间的差异
然后按照差异进行排序

Procedure of Wilcoxon shgned-rank test:

Test value: 120 cm

-11cm +6cm

-1 -2 -3 -6 -7 -9 -10 +8 +5 +4

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

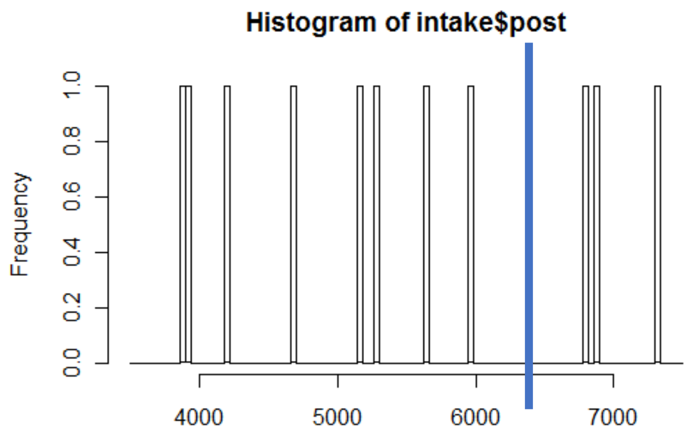
添加符号，-表示低于测试值，+表示高于测试值

- 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 means 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

比较正等级与负等级的平均值，如果样本平均值接近测试值，则正等级和负等级排序的均值不应相差太大

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

Wilcoxon signed-rank test



- Is post-menstrual calorie consumption (*intake$post*) different from 6500 kcal?

conf.int指是否给出相应的置信区间 mu是测试值

- > 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
```

备择假设：真实位置不等于6500

- 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)

假中值类似于中值或均值

- 'pseudomedian' is similar to median or mean
- $P<0.05$

Conclusion:

- reject null hypothesis 拒绝零假设
- post-menstrual calorie consumption is significantly below 6500 kcal

```
> HDR2011$HDI <- as.numeric(HDR2011$HDI)
```
将其从character转换成数字num型

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?

Two-sample Wilcoxon test

双样本t检验的代替方法

= Mann-Whitney test

- alternative to two-sample *t*-test

想要比较男孩和女孩的身高

- We want to compare heights of boys (n=6) and girls (n=4); very small sample!

Similar ranking procedure:

将两个样本混合在一起，从高到低排序，比较两个两本之间的ranks

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

如果男孩女孩身高相似，那么他们的平均身高的排名不应该会有显著差异

10 9 8 7 6 5 4 3 2 1

Two-sample Wilcoxon test

- Example: do !Kung young 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
```

- Mean weight of boys and girls:

```
> tapply(kb$weight, kb$sex, mean, na.rm=T)
     man    woman
6.769861 6.030712
```

- We should run Wilcoxon test (not t-test) because both samples are small

```
> table(kb$sex)
 man woman
  14    20
```

Results:

- W statistic is sum of ranks in first group minus minimum possible value

W统计量是第一组排名之和减去最小可能值

- Difference in location of means = 0.57kg
    - P-value: 0.66
    - 95%5 CI includes 0

- = no significant difference in weight

没有显著差异

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

Exercises:

1) We use Wilcoxon tests when samples are small

a) Open file *zelazo* (with data on walking age in four groups of children); read file description in the ISwR package

b) Compare the groups *active* (children who received active training) and *none* (no training); which test do you use?

c) Now compare *active* and control (*ctr.8w*) groups. Is there a difference?

Note: to call cariables, just write zelazo$active, zelazo$none etc.

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

Which test do you need to use?

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

Matched-pairs Wilcoxon test

配对样本t检验的替代方法

- 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 one-sample and paired-sample t tests to this dataset

But paired-sample Wilcoxon test is the appropriate test due to small sample size (n=11!

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 60 days (time=60) after taking the drug?