

Dealing with assumptions

Chapter 13

Detecting deviations from normality

Previous data / theory

Histograms

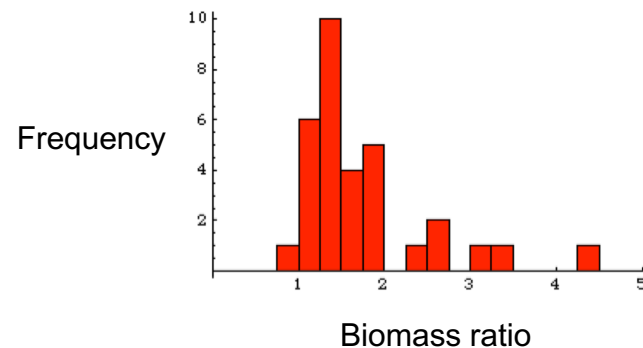
Quantile plots

Shapiro-Wilk test

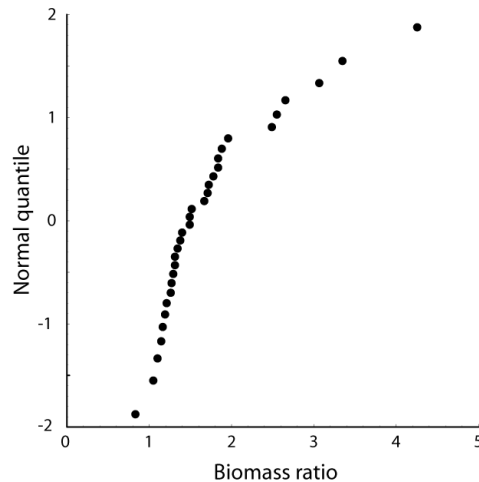
Assumptions of t -tests

- Random sample(s)
- Populations are normally distributed
- (for 2-sample t) Populations have equal variances

Detecting deviations from normality: by histogram

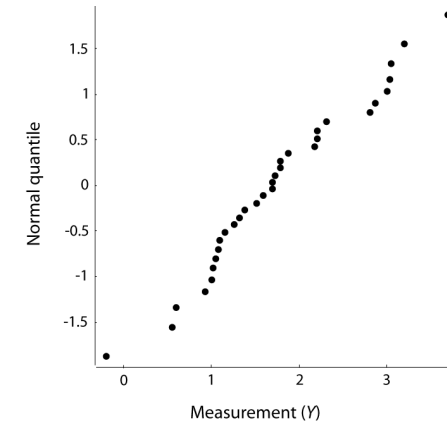


Detecting deviations from normality: by quantile plot



Detecting deviations from normality: by quantile plot

Normal data



Detecting differences from normality: Shapiro-Wilk test

A Shapiro-Wilk test is used to test statistically whether a set of data comes from a normal distribution.

What to do when the assumptions are not true: options

- If the sample sizes are large, sometimes the parametric tests work OK anyway
- Transformations
- Non-parametric tests
- Permutation tests
- Bootstrapping

The normal approximation

Means of large samples are normally distributed.

Therefore, the parametric tests on large samples work relatively well, even for non-normal data.

Rule of thumb: if $n > \sim 50$, the normal approximations may work.

Parametric tests - Unequal variance

Welch's t -test is ideal.

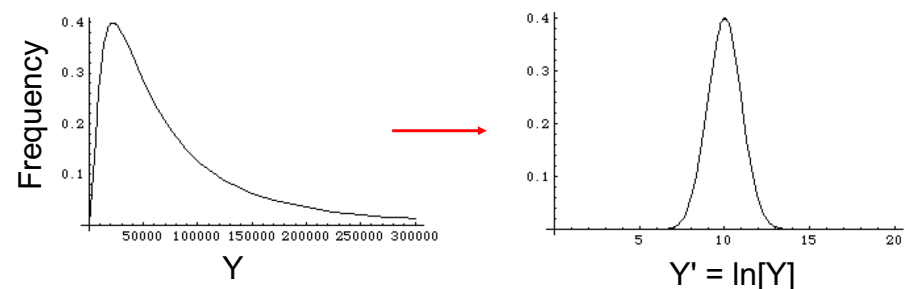
If sample sizes are equal and large, then even a ten-fold difference in variance is *approximately* OK. (But Welch's is still better.)

Data transformations

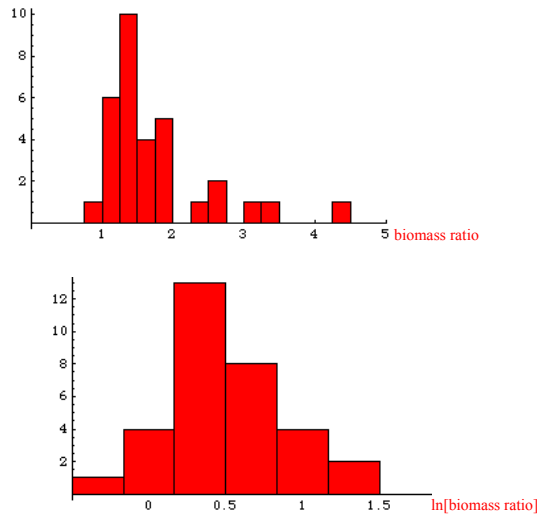
A data transformation changes each data point by some simple mathematical formula.

Log-transformation

$$Y' = \ln[Y]$$



Biomass ratio	ln[Biomass Ratio]
1.34	0.30
1.96	0.67
2.49	0.91
1.27	0.24
1.19	0.18
1.15	0.14
1.29	0.26

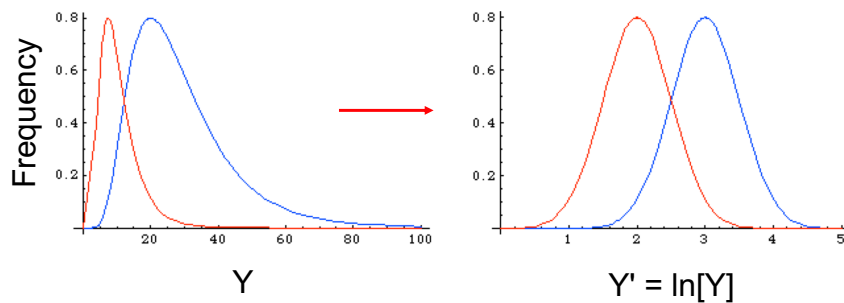


Carry out the test on the transformed data!

The log transformation is often useful when:

- the variable is likely to be the result of multiplication or division of various components.
- the frequency distribution of the data is skewed to the right
- the variance seems to increase as the mean gets larger (in comparisons across groups).

Variance and mean increase together --> try the log-transform



Other transformations

Arcsine: $p' = \arcsin[\sqrt{p}]$

Square-root: $Y' = \sqrt{Y + \frac{1}{2}}$

Reciprocal: $Y' = \frac{1}{Y}$

Example: Confidence interval with log-transformed data

Data:	5	12	1024	12398
Log data:	1.61	2.48	6.93	9.43

$$\bar{Y}' = 5.11 \quad s_{\ln[Y]} = 3.70$$

$$\bar{Y}' \pm \frac{t_{0.05(2),3} s_{\ln[Y]}}{\sqrt{n}} = 5.11 \pm 3.18 \frac{3.70}{\sqrt{4}} = 5.11 \pm 5.88$$

$$-0.993 < \mu_{\ln[Y]} < 10.99$$

Choosing transformations

Must transform each individual in the same way

The transformed values must still carry biological meaning.

You CANNOT keep trying transformations until $P < 0.05!!!$

Valid transformations...

Require the same transformation be applied to each individual

Have one-to-one correspondence to original values

Have a monotonic relationship with the original values (e.g., larger values stay larger)

Non-parametric methods

Assume less about the underlying distributions

Also called "distribution-free"

"Parametric" methods assume a distribution or a parameter

Sign test

Non-parametric test

Compares data from one sample to a constant

Simple: for each data point, record whether individual is above (+) or below (–) the hypothesized constant.

Use a binomial test to compare result to 1/2.

Example: Polygamy and the origin of species

Is polygamy associated with higher or lower speciation rates?

Arnqvist *et al.* (2000) Sexual conflict promotes speciation in insects. *PNAS* 97:10460-10464.

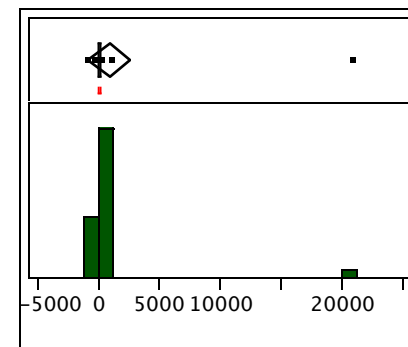
Data:

Order	Family	Multiple mating group	Number of species	Single mating group	Number of species
Beetles	Anobiidae	Ernobius	53	Xestobium	10
	Dermestidae	Dermestes	73	Trogoderma	120
	Elateridae	Agriotes	228	Selatosomus	74
Flies	Muscidae	Coenosia	353	Delia	289
	Cecidomyiidae	Rhopalomyia	157	Mayetiola	30
	Chironomidae	Chironomus	300	Pontomyia	4
	Chironomidae	Stictochironomus	34	Clunio	18
	Drosophilidae	Drosophilidae	3,400	Culicidae	3,500
	Dryomyzidae and Calliphoridae	Dryomyzidae	20	Calliphoridae	1,000
	Tephritidae	Anastrepha	196	Bactrocera	486
	Sciaridae and Bibionidae	Sciaridae	1,750	Bibionidae	660
	Scatophagidae	Scatophaga	55	Musca	63
Mayflies	Siphonuridae	Siphonurus	37	Caenis	115
Homoptera	Psyllidae	Cacopsylla	100	Aonidiella	30
Butterflies and moths	Noctuidae and Psychidae	Noctuidae	21,000	Psychidae	600
	Tortricidae	Choristoneura	37	Epiphyas	40
	Nymphalidae	Eueides (aliphra clade)	7	Eueides (vibilia clade)	5
	Nymphalidae	Heliconius (silvaniform clade)	15	Heliconius (sarasapho clade)	7
	Nymphalidae	Polygonia /	18	Nymphalis	6

Etc....

The differences are not normal

43	-47	154	64	127	296	16
-100	-980	-290	1090	-8	-78	70
20940	-3	2	8	12	227	1
61	1	79	78			



Hypotheses

H₀: The median difference in number of species between singly-mating and multiply-mating insect groups is 0.

H_A: The median difference in number of species between these groups is not 0.

7 out of 25 comparisons are negative

43	-47	154	64	127	296	16
-100	-980	-290	1090	-8	-78	70
20940	-3	2	8	12	227	1
61	1	79	78			

Binomial test on pluses and minuses (compared to $p = 0.5$):

$$\Pr[X \leq 7] = \sum_{i=0}^7 \binom{25}{i} (0.5)^i (0.5)^{25-i} = 0.02164$$

$$P = 2 (0.02164) = 0.043$$

Sign test in R

```
polygamyData$difference =
  polygamyData$SpeciesMultipleMating -
  polygamyData$SpeciesSingleMating

polygamyData$signOfDifference =
  ifelse(polygamyData$difference>0,"Positive", "Negative")

table(polygamyData$signOfDifference)
```

```
Negative Positive
      7      18
```

```
binom.test(7,25)
Exact binomial test
data: 7 and 25
number of successes = 7, number of trials = 25, p-value = 0.04329
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
0.1207167 0.4938768
sample estimates:
probability of success
0.28
```

The sign test has very low power

So it is quite likely to *not* reject a *false* null hypothesis.

Most non-parametric methods use RANKS

Rank each data point in all samples from lowest to highest.

Lowest data point gets rank 1, next lowest gets rank 2, ...

Performing a Mann-Whitney U test

First, rank all individuals from both groups together in order (for example, smallest to largest).

Sum the ranks for all individuals in each group --> R_1 and R_2

Non-parametric test to compare 2 groups

The *Mann-Whitney U test* compares the central tendencies of two groups using ranks.

Calculating the test statistic, U

$$U_1 = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - R_1$$

$$U_2 = n_1 n_2 - U_1$$

U_1 is the number of times an individual from pop. 1 has a lower rank than an individual from pop. 2, out of all pairwise comparisons.

Example: Garter snake resistance to newt toxin



Rough-skinned newt



Comparing snake resistance to TTX (tetrodotoxin)

Locality	Resistance
Benton	0.29
Benton	0.77
Benton	0.96
Benton	0.64
Benton	0.70
Benton	0.99
Benton	0.34
Warrenton	0.17
Warrenton	0.28
Warrenton	0.20
Warrenton	0.20
Warrenton	0.37

This variable is known to be not normally distributed within populations.

Hypotheses

H_0 : The TTX resistance for snakes from Benton is the same as for snakes from Warrenton.

H_A : The TTX resistance for snakes from Benton is different from snakes from Warrenton.

Calculating the ranks

Locality	Resistance	Rank
Benton	0.29	5
Benton	0.77	10
Benton	0.96	11
Benton	0.64	8
Benton	0.70	9
Benton	0.99	12
Benton	0.34	6
Warrenton	0.17	1
Warrenton	0.28	4
Warrenton	0.20	2.5
Warrenton	0.20	2.5
Warrenton	0.37	7

Rank sum for Warrenton: $R=1+4+2.5+2.5+7=17$

Mann-Whitney test in R

(equivalent to Wilcoxon rank sum test)

```
wilcox.test(wholeAnimalResistance ~ locality, data = snakeData)

cannot compute exact p-value with ties
Wilcoxon rank sum test with continuity correction

data: wholeAnimalResistance by locality
W = 33, p-value = 0.01468
alternative hypothesis: true location shift is not equal to 0
```

Permutation tests

Used for hypothesis testing on measures of association

Mixes the real data randomly

Assumptions of Mann-Whitney U test

Both samples are random samples.

Both populations have the same shape of distribution.*

* Only necessary when using Mann-Whitney to compare means.

Permutation tests

1. Variable 1 from an individual is paired with variable 2 data from a randomly chosen individual. This is done for all individuals.
2. The estimate is made on the randomized data.
3. The whole process is repeated numerous times. The distribution of the randomized estimates is the null distribution.

Without replacement

Permutation tests are done without replacement.

In other words, all data points are used exactly once in each permuted data set.

Permutation can be done for any test of association between two variables

Example: Sage crickets

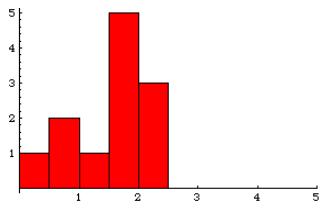


Sage cricket males sometimes offer their hind-wings to females to eat during mating.

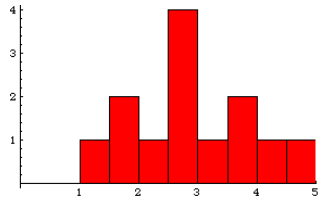
Do females who eat hind-wings wait longer to re-mate?

Waiting time to remating in sage cricket females after initial mating with either a wingless or winged male (presented in $\ln(\text{days})$)

Male wingless	Male winged
0	1.4
0.7	1.6
0.7	1.9
1.4	2.3
1.6	2.6
1.8	2.8
1.9	2.8
1.9	2.8
1.9	3.1
2.2	3.8
2.1	3.9
2.1	4.5
	4.7



ln(Time to remating): First mate had no wings



ln(Time to remating): First mate had intact wings

Problems:
Unequal variance,
non-normal distributions

Real data: $\bar{Y}_1 - \bar{Y}_2 = -1.41$

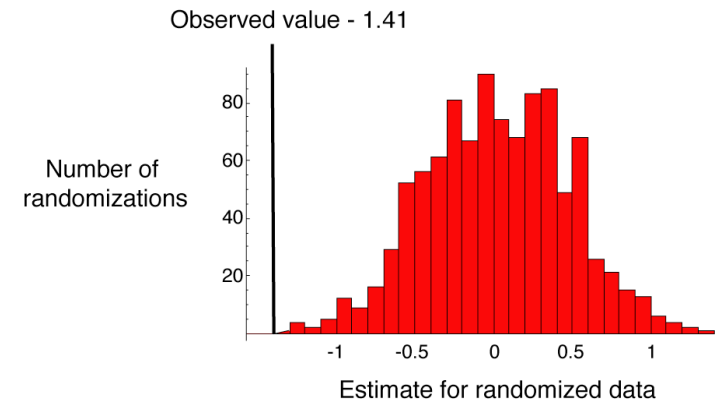
Male wingless	Male winged
0	1.4
0.7	1.6
0.7	1.9
1.4	2.3
1.6	2.6
1.8	2.8
1.9	2.8
1.9	2.8
1.9	3.1
2.2	3.8
2.1	3.9
2.1	4.5
	4.7

Randomized data: $\bar{Y}_1 - \bar{Y}_2 = 0.41$

Male wingless	Male winged
0.7	2.8
2.3	1.9
1.9	2.1
1.8	1.6
3.8	0
1.4	1.4
1.9	2.2
3.9	2.1
4.7	1.6
2.6	4.5
1.9	2.8
2.8	0.7
	3.1

1000 permutations

Note that each data point was
only used once



$P < 0.001$

A permutation approach in R

```
cricketData = read.csv("cricketWingless.csv")

differenceInMeans = function(groupVector, numericVector){
  df = data.frame(groupName = groupVector, y =
    numericVector)

  means = df %>% group_by(groupName) %>%
    summarize(meanOfGroup = mean(y))

  means$meanOfGroup[2] - means$meanOfGroup[1]
}

observedDifference =
  differenceInMeans(cricketData$Treatment,
    cricketData$logDaysToRemating)

observedDifference

[1] -1.413462
```

```
permutationDifferenceInMeans =
  function(groupVector, numericVector){
    n=length(numericVector)
    permutedNumericVector = sample(numericVector,
      size = n, replace=FALSE)
    differenceInMeans(groupVector, permutedNumericVector)
  }

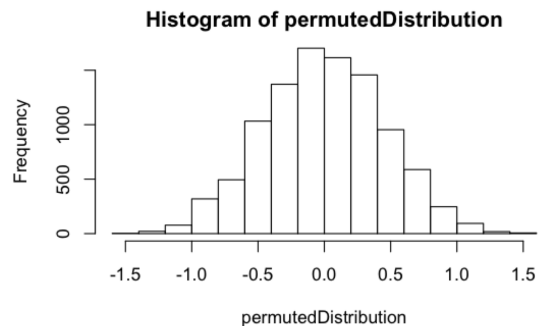
permutationDifferenceInMeans(cricketData$Treatment,
  cricketData$logDaysToRemating)
```

```
[1] -0.6057692
```

Note: this is just one possible answer from the permutation.

```
permutedDistribution = replicate(n=10000,
  permutationDifferenceInMeans(cricketData$Treatment,
    cricketData$logDaysToRemating))

hist(permutedDistribution)
```



```
temp = permutedDistribution<=observedDifference
table(temp)
```

```
temp
FALSE TRUE
9997    3
```

So the P -value for this test is $P = 2 \times 3/10000 = 0.0006$