Stat 302
Statistical Software and Its Applications
Two-Sample Test

Yen-Chi Chen

Department of Statistics, University of Washington

Autumn 2016

An Example: Chicken Weight data – 1

```
> data1 <- chickwts[chickwts$feed=="meatmeal",1]
> data2 <- chickwts[chickwts$feed=="sunflower",1]
> data1
 [1] 325 257 303 315 380 153 263 242 206 344 258
> data2
 [1] 423 340 392 339 341 226 320 295 334 322 297
 318
```

→ The two sample test is to compare these two samples.

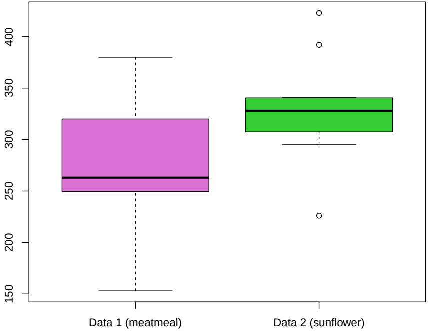
An Example: Chicken Weight data – 2

- Why do we care about comparing these two samples?
- If you are a scientist, you may want to know if the `feed` for chicken affects their growth (`weight`).
- If you are a businessman, you may be interested in if the `feed` changes the weight of chicken (so that you can make money by using the best `feed`).
- In many situations, we would like to see if the two samples are different or not.
- If the `feed` and `weight` are independent, then the distributions of the two samples will be the same.
- Today we will talk about two classes of approaches: visual comparison and quantitative comparison.

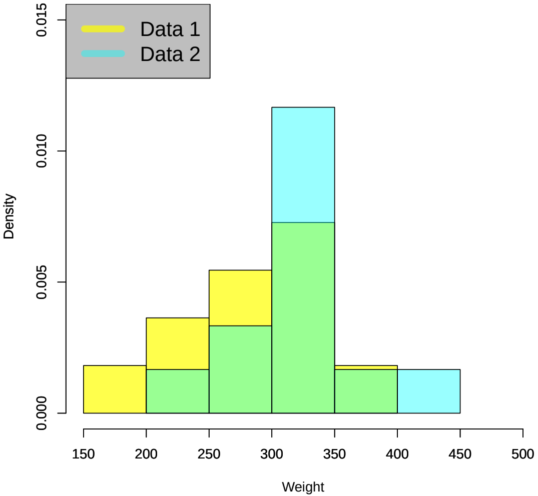
Visual Comparison: Boxplot

Showing boxplot for both samples is one way to compare them.

```
> boxplot(data1,data2, col=c("orchid","limegreen"),
+             names=c("Data 1 (meatmeal)",
+                     "Data 2 (sunflower)"))
```



Overlapping histograms is another approach.



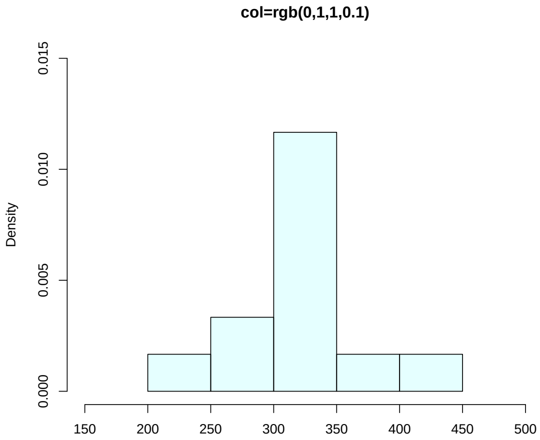
Visual Comparison: Histogram – 2

```
> hist(data1, col=rgb(1,1,0,0.7), ylim=c(0,0.015),
+       xlim=c(150,500), probability=T,
+       main="", xlab="Weight")
> par(new=T)
> hist(data2, col=rgb(0,1,1,0.4), ylim=c(0,0.015),
+       xlim=c(150,500), probability=T,
+       main="", xlab="")
> legend("topleft", c("Data 1","Data 2"),
+          col=c(rgb(1,1,0,0.7),rgb(0,1,1,0.4)),
+          lwd=8, cex=1.5, bg="gray")
```

- `col`: we need to use transparent color.
- `probability`: we need it to be `T` because two samples may have different sample size.
- `par(new=T)`: the next plot will be overlapped with the previous plot.

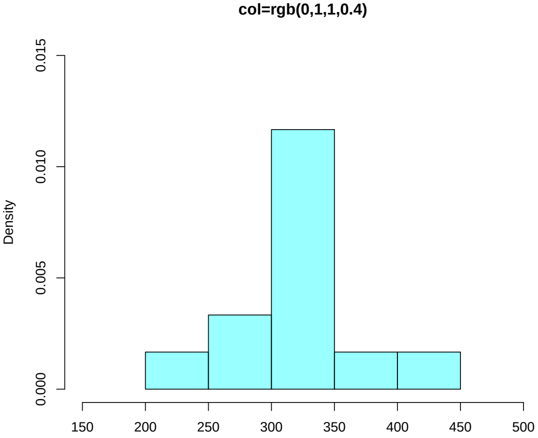
Transparent color – 1

```
> hist(data2, col=rgb(0,1,1,0.1), ylim=c(0,0.015),
+       xlim=c(150,500), probability=T,
+       main="col=rgb(0,1,1,0.1)", xlab="")
```



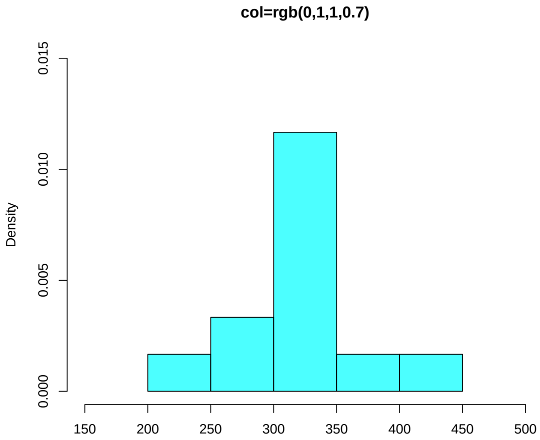
Transparent color – 2

```
> hist(data2, col=rgb(0,1,1,0.4), ylim=c(0,0.015),
+       xlim=c(150,500), probability=T,
+       main="col=rgb(0,1,1,0.4)", xlab="")
```



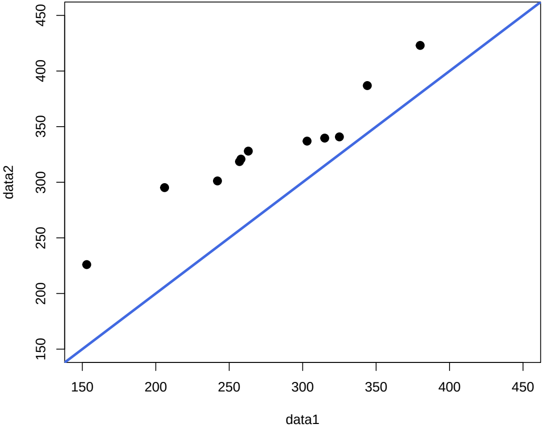
Transparent color – 3

```
> hist(data2, col=rgb(0,1,1,0.7), ylim=c(0,0.015),
+       xlim=c(150,500), probability=T,
+       main="col=rgb(0,1,1,0.7)", xlab="")
```



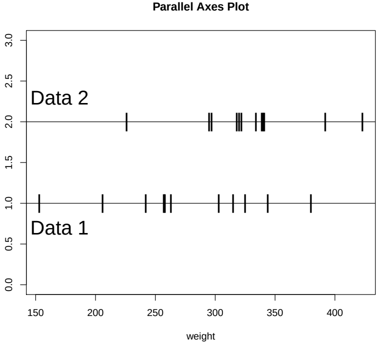
Visual Comparison: QQ plot

```
> qqplot(data1, data2, xlim=c(150, 450),
+          ylim=c(150,450))
> abline(a=0,b=1, lwd=3, col="royalblue")
```



Visual Comparison: Parallel Axes plot

```
> plot(x=c(data1,data2), y=c(rep(1, length(data1),)
+       rep(2, length(data2))), pch="|",
+       ylim=c(0,3), cex=2, ylab="", xlab="weight",
+       main="Parallel Axes Plot")
> text(x=170,y=0.7, labels="Data 1", cex=2)
> text(x=170,y=2.3, labels="Data 2", cex=2)
> abline(h=1);abline(h=2)
```



Quantitative Comparison: Hypothesis Test

- In many cases, visual comparison is not enough.
- We want some quantitative way to compare two samples.
- One quantitative approach is to frame the problem using *the hypothesis test*.
- In English: we want to know *if the two samples are from the same distribution*.
- In Statistics, the above question can be viewed as testing the following null hypothesis:

$$H_0 : \text{two samples are from the same distribution.}$$

- Let P_1 be the population distribution of data 1 and P_2 be the population distribution of data 2.
- Then the above H_0 is equivalent to

$$H_0 : P_1 = P_2.$$

Hypothesis Test

- The goal is to test
$$H_0 : P_1 = P_2.$$
- There are several methods to test the above procedure.
- These methods can be divided into two groups: parametric methods and nonparametric methods.
- Parametric methods: we use some parameters of the distribution to carry out the test.
- Examples of parametric methods: mean test and variance test.
- Nonparametric methods: we directly use the entire distribution to do testing.
- Examples of nonparametric methods: KS-test and rank test.

Parametric Method: Mean Test – 1

- Because
$$H_0 : P_1 = P_2$$
implies $\mu_1 = \mu_2$ (μ_i is the mean of P_i), the mean test is to test
$$H_0 : \mu_1 = \mu_2, .$$
- Testing $\mu_1 = \mu_2$ is equivalent to testing
$$H_0 : \mu_1 - \mu_2 = 0.$$
- So the test statistics is to use the difference between sample means $\bar{X}_1$ and $\bar{X}_2$ and rescale it by the variance.

Parametric Method: Mean Test – 2

- Assume the sample 1 consists of IID $X_{1,1}, \cdots, X_{1,n}$ and the sample 2 consists of IID $X_{2,1}, \cdots, X_{2,m}$ and sample 1 and sample 2 are independent from each other.
- Then the sample means have variance

$$\mathrm{Var}(\bar{X}_1) = \frac{\sigma_1^2}{n}, \quad \mathrm{Var}(\bar{X}_2) = \frac{\sigma_2^2}{m},$$

  where σ_1^2 and σ_2^2 are the variance of P_1 and P_2.
- Thus, the quantity $\bar{X}_1 - \bar{X}_2$ has variance $\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}$ (why?).
- Because we do not know σ_1^2 and σ_2^2 in practice, we will replace them by the sample variance S_1^2 and S_2^2.
- Thus, our final test statistics is

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n} + \frac{S_2^2}{m}}}.$$

- T will follow asymptotically a standard normal distribution (think about why) so we can compare T to the standard normal to obtain a p-value.

Parametric Method: Mean Test – 3

- Test statistics is

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n} + \frac{S_2^2}{m}}}.$$

- We called this approach Z-test because we use the feature that the asymptotic distribution is a standard normal.

```
> mean1 <- mean(data1)
> mean2 <- mean(data2)
> sd1 <- sd(data1)/sqrt(length(data1))
> sd2 <- sd(data2)/sqrt(length(data2))
> Test.stat <- (mean1-mean2)/sqrt(sd1^2+sd2^2)
> 2*(1-pnorm(abs(Test.stat)))
[1] 0.03105238
> 2*(pnorm(-abs(Test.stat)))
[1] 0.03105238
```

- So the p-value is 0.03105238.

Parametric Method: Mean Test – 4

- Another approach is to use the T-test.
- If we assume P_1 and P_2 are from the same normal distribution, then the test statistics

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n} + \frac{S_2^2}{m}}}$$

  follows a T-distribution with a complicated degree of freedom:

$$\nu = \frac{(S_1^2/n + S_2^2/m)^2}{\frac{(s_1^2/n)^2}{n-1} + \frac{(s_2^2/m)^2}{m-1}}.$$

- In R, there is a built-in function `t.test()` that allows us to T-test.

Parametric Method: Mean Test – 5

```
> t.test(data1,data2)

        Welch Two Sample t-test

data:  data1 and data2
t = -2.1564, df = 18.535, p-value = 0.04441
alternative hypothesis: true difference in means
is not equal to 0

95 percent confidence interval:
 -102.572435   -1.442716
sample estimates:
mean of x mean of y
 276.9091  328.9167
```

Parametric Method: Mean Test – 6

- So there are two approaches for testing the mean: Z-test and T-test.
- There is no definitely which test is better than the others because they rely on different assumptions.
- The Z-test requires very weak assumption on data–we do not need to assume the true distribution is a normal distribution.
- But the Z-test only *works asymptotically*; namely, it works when sample size is large enough.
- The T-test requires a strong assumption: the distribution is a normal distribution.
- However, if the samples are from normal distributions, T-test *works regardless of the sample size*.

Parametric Method: Variance Test – 1

- Because
$$H_0 : P_1 = P_2$$
implies $\sigma_1^2 = \sigma_2^2$, the variance test is to test
$$H_0 : \sigma_1^2 = \sigma_2^2.$$
- The null hypothesis $H_0 : \sigma_1^2 = \sigma_2^2$ is equivalent to
$$H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1.$$
- So the test statistics is to use the ratio between sample variance $\frac{\bar{S}_1^2}{\bar{S}_2^2}$.
- When the two samples are from the same normal distribution, the test statistics $\frac{\bar{S}_1^2}{\bar{S}_2^2}$ follows a distribution called F-distribution.
- In R, you can use the command `var.test()` to carry out variance test.

Parametric Method: Variance Test – 2

```
> var.test(data1,data2)

        F test to compare two variances

data:  data1 and data2
F = 1.7661, num df = 10, denom df = 11,
p-value = 0.3645
alternative hypothesis: true ratio of variances
is not equal to 1

95 percent confidence interval:
 0.5009206 6.4725366
sample estimates:
ratio of variances
          1.766081
```

Nonparametric Method: KS-test – 1

- The nonparametric test directly test $H_0 : P_1 = P_2$.
- The KS-test (Kolmogorov-Smirnov test) is a classical approach in nonparametric two-sample test.
- Given $X_{1,1}, \cdots, X_{1,n}$ IID from P_1, we can estimate P_1 by the *empirical distribution function (EDF)*:

$$\hat{P}_1(t) = \frac{1}{n} \sum_{i=1}^{n} I(X_{1,i} \leq t),$$

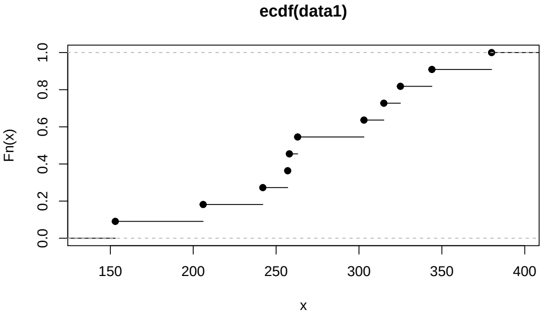
  where $I(x)$ is the indicator function.
- $\hat{P}_1(t)$ is the ratio of data points whose value is below t.
- Note: the definition of the distribution P_1 is

$$P_1(t) = P(X_{1,i} \leq t).$$

- The EDF can be computed using function *ecdf*().

Nonparametric Method: KS-Test – 2

```
> ecdf(data1)
Empirical CDF
Call: ecdf(data1)
 x[1:11] =    153,    206,    242,  ...,    344,
380
>
> plot(ecdf(data1))
```



- The KS-test is to use the following test statistics:

$$K = \sup_t |\hat{P}_1(t) - \hat{P}_2(t)|.$$

- After rescaling, the test statistics K has a known limiting distribution called the *Kolmogorov distribution*.
- An appealing feature is that the Kolmogorov distribution does not depend on the true distribution P_1 and P_2.
- In R, we use the command `ks.test()` to carry out the KS-test.

```
> ks.test(data1,data2)

        Two-sample Kolmogorov-Smirnov test

data:  data1 and data2
D = 0.47727, p-value = 0.1085
alternative hypothesis: two-sided
```

Nonparametric Method: Rank Test – 1

- Now we introduce another nonparametric test: rank test.
- This test is also known as the Wilcoxon Rank Sum test or Mann-Whitney test.
- Recalled that we want to test $H_0 : P_1 = P_2$.
- The rank test is to first pull the two samples together, computing the rank of each data point.
- Then use the sum of the rank of the data points from sample 1 as a test statistics.
- Under H_0, the two distributions are the same so the rank of data points from sample 1 should be uniformly distributed within $\{1, 2, \cdots, n + m\}$.
- In R, we will use the command `wilcox.test()`.

Nonparametric Method: Rank Test – 2

```
> data_all <- c(data1,data2)
> idx_all <- c(rep(1, length(data1)),
+              rep(2, length(data2)))
> rank_idx <- rbind(rank(data_all)[order(data_all)],
+             idx_all[order(data_all)])
> row.names(rank_idx) <- c("Rank", "Sample")
> rank_idx
       [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9]
Rank      1    2    3    4    5    6    7    8    9
Sample    1    1    2    1    1    1    1    2    2
       [,10] [,11] [,12] [,13] [,14] [,15] [,16] [,17]
Rank      10    11    12    13    14    15    16    17
Sample     1     1     2     2     2     1     2     2
       [,18] [,19] [,20] [,21] [,22] [,23]
Rank      18    19    20    21    22    23
Sample     2     2     1     1     2     2
```

Nonparametric Method: Rank Test – 3

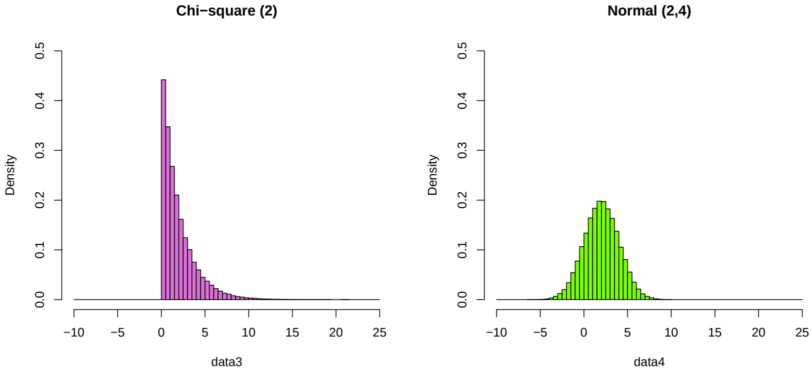
```
> wilcox.test(data1,data2)

        Wilcoxon rank sum test

data:   data1 and data2
W = 36, p-value = 0.06882

alternative hypothesis: true location shift
is not equal to 0
```

Nonparametric Method: Comments

- Nonparametric methods require a weaker assumption on the distribution.
- However, the power of the nonparametric tests is generally lower than the parametric approach.
- Namely, nonparametric tests tend to have a higher p-value than the parametric approach when H_0 is false and the parametric assumption is reasonable.
- In additional to the KS-test and rank test, there are many other nonparametric tests.
- For instance, we can use the difference in histogram to test two samples.
- Nonparametric two-sample test is still a very popular research field in both statistics and machine learning.

Case study: Chi-Square versus Normal – 1

- Here we consider generating data from two distributions: a chi-square distribution and a Normal distribution.
- We compare the chi-square distribution with degree of freedom 2 and the Normal distribution with mean 2 variance 4.



Case study: Chi-Square versus Normal – 2

- The two distributions apparently look very different from each other.
- Now we generate 200 data points from each of these two distributions and compare them.

```
> set.seed(1)
> data3<- rchisq(n=200, df=2)
> data4<- rnorm(n=200, mean = 2, sd=2)
```

Let's first try *Z*-test:

```
> mean3 <- mean(data3)
> mean4 <- mean(data4)
> sd3 <- sd(data3)/sqrt(length(data3))
> sd4 <- sd(data4)/sqrt(length(data4))
> Test.stat <- (mean3-mean4)/sqrt(sd3^2+sd4^2)
>
> 2*(1-pnorm(abs(Test.stat)))
[1] 0.6430986
```

$\rightarrow$ Not significant.

Case study: Chi-Square versus Normal – 4

Now we try T-test:

```
> t.test(data3,data4)

        Welch Two Sample t-test

data:  data3 and data4
t = 0.46337, df = 396.18, p-value = 0.6434
alternative hypothesis: true difference in means
is not equal to 0

95 percent confidence interval:
 -0.3196255  0.5167574
sample estimates:
mean of x mean of y
 2.047846  1.949280
```

→ Also not significant.

Case study: Chi-Square versus Normal – 5

Now we try variance test:

```
> var.test(data3,data4)
        F test to compare two variances
data:  data3 and data4
F = 1.1452, num df = 199, denom df = 199,
p-value = 0.3397

alternative hypothesis: true ratio of variances
is not equal to 1
95 percent confidence interval:
 0.8666766 1.5132484
sample estimates:
ratio of variances
          1.145206
```

$\rightarrow$ Still... not significant.

Case study: Chi-Square versus Normal – 6

Now we try KS-test:

```
> ks.test(data3,data4)

        Two-sample Kolmogorov-Smirnov test

data:  data3 and data4
D = 0.175, p-value = 0.004375
alternative hypothesis: two-sided
```

→ Now we get a significant result!

Case study: Chi-Square versus Normal – 7

How aboutrank test:

```
> wilcox.test(data3,data4)

Wilcoxon rank sum test with continuity correction

data:  data3 and data4
W = 18990, p-value = 0.3826
alternative hypothesis: true location shift
is not equal to 0
```

→ Does not work.

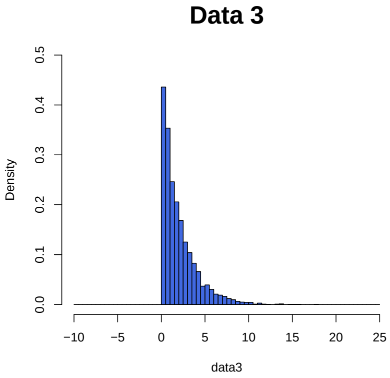
Case study: Chi-Square versus Normal – 8

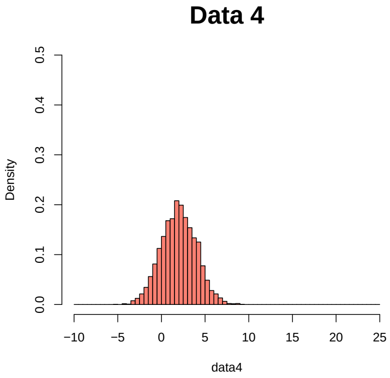
- The reason why most tests fail is because the two distributions have the same mean and the variance!
- This is the power of a nonparametric test; the KS-test is still capable of detecting the difference even when the mean and variance are the same in both sample.

Now when we increase the sample size:

```
> data3<- rchisq(n=5000, df=2)
> data4<- rnorm(n=5000, mean = 2, sd=2)
>
> t.test(data3,data4)$p.value
[1] 0.1397539
> var.test(data3,data4)$p.value
[1] 0.4311391
> ks.test(data3,data4)$p.value
[1] 0
> wilcox.test(data3,data4)$p.value
[1] 1.431667e-06
```

→ The nonparametric tests work but the parametric tests still fail.


Data 3
Density
0.0
0.1
0.2
0.3
0.4
0.5
−10
−5
0
5
10
15
20
25
data3


Data 4
Density
0.0
0.1
0.2
0.3
0.4
0.5
−10
−5
0
5
10
15
20
25
data4

In-class Exercises

- Go back to `chickwts` dataset.
- Now try to compare the weight of group whose `feed` is `casein` versus `horsebean`.
- Use visual comparison to compare them.
- Use quantitative comparison to test if the two samples are significantly different from each other.