

BIOC461

Research Design and Analysis in Biochemistry

Lecture 2

A/Prof Mik Black
Department of Biochemistry
24 March 2020

Wilcoxon–Mann–Whitney Test (WMW)

- t-test works well for two independent samples when the two populations have normal (or near-normal) distributions.
- Serious problems can occur when the distributions have heavy tails or are skewed.
- Wilcoxon-Mann-Whitney test does not depend on the assumption of normality.
 - H_0 : the two populations have same distribution
 - H_A : the two populations have different distributions
- Calculate a test statistic U_s .

Example

- Counting the number of seeds produced by two related species of plant:
 - Do the number of seeds produced by the two plant species have the same distribution (at $\alpha = 0.05$)?
- No parameters to define:
 - H_0 : the two species have the same distribution of seed number
 - H_A : the two species have different distribution of seed number

Example (continued)

- Data should be ordered, if not sort them:

Species 1: 19, 23, 25, 28, 28, 34 (n1 = 6)

Species 2: 14, 18, 19, 20, 25 (n2 = 5)

- Calculate K_1 : for each observation in sample 1, count the number of observations in sample 2 that are smaller than it; add them up (ties count $\frac{1}{2}$):

$$K_1 = 2 \frac{1}{2} + 4 + 4 \frac{1}{2} + 5 + 5 + 5 = 26$$

- K_2 : for each observation in sample 2, count the number of those in sample 1 that are smaller than it (ties count $\frac{1}{2}$):

$$K_2 = 0 + 0 + \frac{1}{2} + 1 + 2 \frac{1}{2} = 4$$

Example (continued)

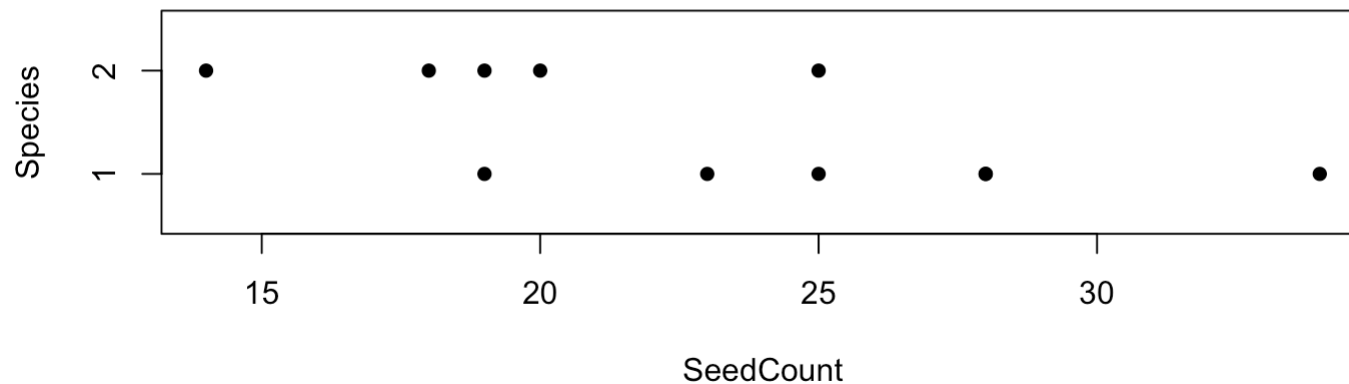
- Test statistic: $U_s = (\text{the larger of } K_1 \text{ and } K_2) = \max(K_1, K_2) = 26$.
- $U_s = \max(K_1, K_2)$ follows the Mann-Whitney null distribution.
- Test at level $\alpha = 0.05$; critical value is $U_{critical} = 27$.
 - $U_s = 26 < 27 = U_{critical}$ (so P-value > 0.05)
 - *Do not* reject H_0 .
- These data do not provide evidence at the 0.05 significance level that the two plant species have different seed number distributions.

Wilcoxon test in R

```
seeds<-read.csv( 'DataSets2/seedcounts.csv' )
```

```
View(seeds)
```

```
stripchart(SeedCount ~ Species, data=seeds, pch=16,  
           ylab="Species", ylim=c(0.5,2.5))
```



Wilcoxon test in R

```
wilcox.test(SeedCount ~ Species, data=seeds)
```

```
## Warning in wilcox.test.default(x = c(28L, 23L, 19L, 28L, 25L, 34L), y =  
## c(19L, : cannot compute exact p-value with ties
```

```
##
```

```
## Wilcoxon rank sum test with continuity correction
```

```
##
```

```
## data: SeedCount by Species
```

```
## W = 26, p-value = 0.05358
```

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

Paired Designs

- Sometimes observations occur in pairs:
 - Two observations on the same individual (two days, two sides, before/after...)
 - observing two groups over time
- Examples:
 - subjects occur naturally in pairs, such as identical twins
 - subjects matched in pairs for similar age, sex, profession, disease status, etc (to match extraneous variables)
 - same subject measured on two occasions

Paired sample t-test

- A combination of two things we already know - two-sample and one sample t-tests (also Confidence Intervals):
 1. Recognize that it is a paired design.
 2. Do hypothesis testing steps like a two independent sample t-test
 3. Do t calculations like a 1-sample t-test but for differences of measurements.

Example

- Have 12 fruit flies from 6 different lines (**2 of each**) grown in vials at **two different temperatures**.
- After a specified time for development, count the number of bristles on each fly.
- We think that bristle number has approximately a normal distribution, but flies from the same line should have similar bristle numbers.
- We will look at the **difference** in bristle number **between two flies of the same line**.

Example (continued)

- 6 “blocks” (lines).
- 2 treatments = temperature (cold/warm)
- Have 12 observations but they occur in 6 pairs, so $n = 6$.

| Line | Cold | Warm | difference C – W |
|------|------|------|-------------------------------|
| 1 | 24 | 25 | –1 |
| 2 | 41 | 43 | –2 |
| 3 | 44 | 46 | –2 |
| 4 | 30 | 31 | –1 |
| 5 | 28 | 31 | –3 |
| 6 | 43 | 43 | 0 |
| mean | 35 | 36.5 | –1.5 |
| s | | | 1.049 |
| SE | | | $\frac{s}{\sqrt{6}} = -0.428$ |

- Does temperature affect the number of bristles on fruit flies?

Example (continued)

- Let 1 and 2 be the mean bristle number for flies grown at the colder and warmer temperatures, respectively:
 - $H_0: \mu_1 = \mu_2$ (i.e., $\mu_1 - \mu_2 = 0$): the mean bristle number is the same for both temperatures
 - $H_A: \mu_1 \neq \mu_2$ (i.e., $\mu_1 - \mu_2 \neq 0$): the mean bristle number is different at the two temperatures
- Use a non-directional paired sample t-test.
 - Test statistic: $t_s = \frac{\bar{y}_{diff}}{SE_{diff}}$
 - Use t-distribution to find p-value.
- Will reject H_0 if $p < \alpha = 0.05$, otherwise will not reject H_0 .

Example (continued)

- Do a one-sample t-test using the difference column only:
 - $\bar{y}_{diff} = 35 - 36 = -1.5$
 - $S_{diff} = 1.049$
 - $SE_{diff} = S_{diff}/\sqrt{n} = 1.049/\sqrt{6} = 0.428$
 - $t_s = -1.5/0.428 = -3.503$
 - Gives $p=0.017$ (see next page), so reject H_0 .
- This study provides evidence at the 0.05 significance level that flies have different mean bristle numbers at the two temperatures.

Paired sample t-test in R

```
flies<-read.csv('DataSets2/flytemp.csv')
```

```
View(flies)
```

```
t.test(flies$Cold, flies$Warm, paired=TRUE)
```

```
##
```

```
## Paired t-test
```

```
##
```

```
## data: flies$Cold and flies$Warm
```

```
## t = -3.5032, df = 5, p-value = 0.01722
```

```
## alternative hypothesis: true difference in means is not equal to 0
```

```
## 95 percent confidence interval:
```

```
## -2.6006574 -0.3993426
```

```
## sample estimates:
```

```
## mean of the differences
```

```
## -1.5
```

Paired tests

- How to know whether to use paired or unpaired t-test to analyze a given data set?
 - Easy: Does each data point in one sample naturally correspond to one specific data point in the other sample? If so, use paired t-test.
- How to know when to use a paired design in your experiment?
 - Harder: Do you expect the extraneous variables to increase variation and want to match on these variables, or pairing is dictated by the problem (e.g. repeated measurements)?
- If so use paired design.

Assumptions

- Paired-sample t-test is based on the idea that the *differences* are approximately normal, so may want to plot the histogram.
- This will automatically be true if the individual groups are both normally distributed.

Sign test

- What if the data are not normally distributed?
 - Recall: for two independent samples did Wilcoxon-Mann-Whitney.
- For paired samples we can do the **sign test**.
- Look at the **sign** of the *difference between each pair of observations*.
- If the two treatments are the same, the sign is equally likely to be positive as negative. That is, if we take a pair at random (from the population of pairs), whether the first or second of the pair is larger is like flipping a coin.

Sign test

- Let π = probability that, for any pair from the population, the first will be larger.
 - $H_0: \pi = 0.5$; the two treatments are the same in their effect.
 - $H_A: \pi \neq 0.5$ (this is the non-directional alternative): the two treatments are different in their effect.
1. For each pair of observations, note whether $y_1 - y_2$ is positive (+) or negative (-), that is, + if y_1 is bigger, - if y_2 is bigger
 2. count the number of + (= N_+) and - (= N_-) (don't count zeros).
 3. test statistic B_s = larger of N_+ and N_- .
 4. Reject H_0 if $B_s \geq$ critical value.

Example: skin grafts

- Skin grafts applied to *both sides of body* in 11 recipients.
- One graft has good HLA antigen system match with recipient; the other does not.
- This is paired because we have two observations from each person.
- Observe time to rejection of skin graft (not normally distributed so don't use t-test.)

Does a good HLA match increase graft survival time?

- Let π = probability that good HLA match survives longer in any individual
 - $H_0: \pi = 0.5$; HLA match and non-match are equally likely to survive longer.
 - $H_A: \pi > 0.5$; HLA match is more likely to survive longer (directional alternative).
- Use a sign test.
 - $B_s = N_+$ has a binomial distribution ($n=11, \pi=0.5$) under H_0 . If non-directional, B_s would be $\max(N_+, N_-)$.
- Test at level $\alpha = 0.05$; critical value is 9 (from computer).
- Will reject H_0 if $B_s \geq 9$; if $B_s < 9$ we will not reject.

Example (continued)

- Data: survival times (days)

| | | | | | | | | | | | |
|------|----|----|----|----|----|----|----|----|----|----|----|
| good | 37 | 19 | 57 | 93 | 16 | 23 | 20 | 63 | 29 | 60 | 18 |
| poor | 29 | 13 | 15 | 26 | 11 | 18 | 26 | 43 | 18 | 42 | 19 |
| sign | + | + | + | + | + | + | - | + | + | + | - |

- $n = 11$
- $N_+ = 9, N_- = 2; B_s = N_+ = 9; B_{0.05} = 9$
- So, reject H_0 .
- This study provides evidence ($P = 0.033$) that a good HLA match survives longer than a poor match.

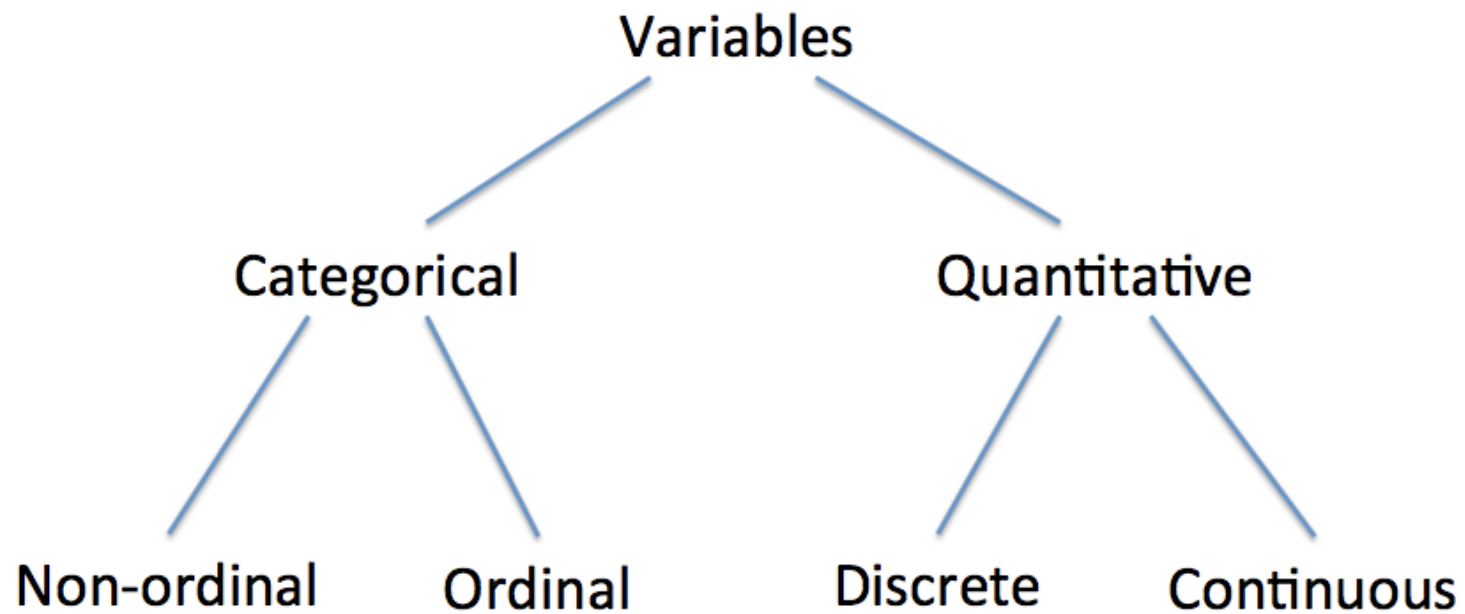
Sign test in R

```
skin <- read.csv('DataSets2/skingraft.csv')

## Perform Sign Test using binom.test function (1-sided alternative)
## Input the test statistic (Bs=9), and the number of pairs (n=11)
binom.test(9, 11, alternative="greater")

##
## Exact binomial test
##
## data: 9 and 11
## number of successes = 9, number of trials = 11, p-value = 0.03271
## alternative hypothesis: true probability of success is greater than 0.5
## 95 percent confidence interval:
## 0.5299132 1.0000000
## sample estimates:
## probability of success
## 0.8181818
```

Types of Variables



Types of Variables

- Quantitative:
 - Outcome is a number
 - height or concentration (continuous)
 - number of flowers on a plant (discrete)
- Categorical:
 - outcomes fall into categories we can list
 - peas: round & yellow, round & green, wrinkled & yellow, wrinkled & green (not ordinal)
 - choices on a survey: never, rarely, occasionally, often, always (ordinal)

Categorical data

- In order to analyze categorical data we start by counting the number of observations in each category.
- If there are only two categories,
 - the number in one category has a binomial distribution (if the observations in the sample are independent).
 - If there are more than two categories,
 - you can focus on one category and group all the others together; still binomial
 - or, more often, you define different π 's for the different categories (π_1, π_2, \dots)

Binomial distribution

- π is the (unknown) probability of “success”, say in category 1.
- n is the number of observations.
- Observe $y = \#$ in category 1.
- The proportion in category 1 is $\hat{\pi} = y/n$.
- y has a binomial (n, π) sampling distribution,
- If n and $n(1-\pi)$ are not too small, this is close to a normal distribution with mean and SD:

$$\mu = n\pi, \quad \sigma = \sqrt{n\pi(1 - \pi)}$$

- This normal approximation is used both explicitly (for confidence intervals) and implicitly (for testing).

Chi-squared distribution

- To test hypotheses, we will use a distribution called a χ^2 distribution (chi-squared).
- Facts about the χ^2 distribution
 - If you have k random variables y_1, \dots, y_k , which are all independent and $N(0,1)$, if you square them all and add them up, the resulting sum has a χ_k^2 distribution. ("a chi-squared distribution with k degrees of freedom").
 - χ^2 random variables only take on positive values.
- We have several different tests which all use the χ^2 distribution to determine critical values, however, the hypotheses are different.

Contingency tables

- “2x2” means 2 rows and 2 columns in a table: categorical data with 4 groups that are related in pairs.
- Two main contexts (sometimes blurred):
 - Two independent random samples; one binomial variable observed in each sample
 - One sample; observe two different binomial variables on each sample
- Examples of context 1:
 1. samples are “drug” and “placebo” (or any two treatments); observed variable “improve” or “don’t improve”.
 2. samples are “male” and “female” (any two groups we set up to compare); observed variable “eye color” brown or blue.

Contingency tables

- Examples of context 2:
 - observe whether people smoke, exercise
 - observe response to drug for patients in a randomized clinical trial. Four Categories, observations in 2x2 table:

| Observed | | treatment | | Total |
|----------|---------------|-----------|---------|-------|
| | | Drug | Placebo | |
| Outcome | Improve | 15 | 4 | 19 |
| | Don't Improve | 11 | 17 | 28 |
| Total | | 26 | 21 | 47 |

- Test for independence of row and column variables.

Association testing: example

| Observed | | treatment | | Total |
|----------|---------------|-----------|---------|-------|
| | | Drug | Placebo | |
| Outcome | Improve | 15 | 4 | 19 |
| | Don't Improve | 11 | 17 | 28 |
| Total | | 26 | 21 | 47 |

- π_1 = probability patient improves if they take the drug
- π_2 = probability patient improves if they take the placebo
 - $H_0: \pi_1 = \pi_2$
 - $H_A: \pi_1 \neq \pi_2$ (or $\pi_1 > \pi_2$)
- H_0 is that the two probabilities are the same (like $\mu_1 = \mu_2$ in a t-test), which can be phrased in terms of independence.

Example (continued)

- H_A is that the probabilities are different, which can be phrased in terms of a *lack of independence*, which is called “*association*”:
 - $\hat{\pi}_1 = \# \text{ who improve with drug} / \# \text{ who take drug}$
 $= 15/26 = 0.58$
 - $\hat{\pi}_2 = \# \text{ who improve with placebo} / \# \text{ who take placebo}$
 $= 4/21 = 0.19$

Example (continued)

- What values would we expect under H_0 ?
 - Total who improve is 19, so under H_0 estimate $19/47 = 40.4\%$ to be the proportion who improve overall.
 - 26 patients took the drug: under H_0 expect 40.4% of them to improve = $(26 \times 0.404) = 10.5 = (26 \times 19)/47$. Similarly, expect number who improve under placebo to be $(21 \times 19)/47 = 8.5$.
 - For the next row, expect $(26 \times 28)/47 = 15.5$ don't improve with drug and $(21 \times 28)/47 = 12.5$ don't improve with placebo.
- Can place these "Expecteds" in a table similar to the observed values. Note that the row and column totals are the same as for observed.

Example (continued)

| Expected | | Treatment | | Total |
|----------|---------------|-----------|---------|-------|
| | | Drug | Placebo | |
| Outcome | Improve | 10.5 | 8.5 | 19 |
| | Don't Improve | 15.5 | 12.5 | 28 |
| Total | | 26 | 21 | 47 |

- In general, for each category:
 - $E = (\text{row total})(\text{column total})/(\text{grand total})$
- This is the common way to organize the E's and the O's:

| Observed (Expected) | | Treatment | | Total |
|---------------------|---------------|-----------|-----------|-------|
| | | Drug | Placebo | |
| Outcome | Improve | 15 (10.5) | 4 (8.5) | 19 |
| | Don't Improve | 11 (15.5) | 17 (12.5) | 28 |
| Total | | 26 | 21 | 47 |

Example - questions

- Are patients who take the drug more likely to improve than those who take the placebo?
 - π_1 = probability patient improves if they take the drug
 - π_2 = probability patient improves if they take the placebo
 - $H_0: \pi_1 = \pi_2$; the probability of improving is the same whether drug or placebo is taken (treatment and outcome are independent)
 - $H_A: \pi_1 > \pi_2$; the probability of improving is greater if the drug is taken than if the placebo is taken
- This example is directional; if non-directional, H_A : outcome and treatment are not independent.

Example (continued)

- Use a χ^2 test of independence
 - $X_s^2 = \sum (O - E)^2 / E$ has a χ_1^2 distribution under H_0 .
 - Calculate p-value, and reject H_0 if $p < \alpha$.
- Since H_A is directional, must do the following step, but it's a good idea whether directional or not.
 - $\hat{\pi}_1 = \frac{\# \text{ who improve with drug}}{\# \text{ who take drug}} = \frac{15}{26} = 0.58$
 - $\hat{\pi}_2 = \frac{\# \text{ who improve with placebo}}{\# \text{ who take placebo}} = \frac{4}{21} = 0.19$
- Check: $0.58 > 0.19$, so $\hat{\pi}_1 > \hat{\pi}_2$ is in the same direction as H_A .

Example (continued)

- $X_s^2 = (15-10.5)^2/10.5 + (4-8.5)^2/8.5 + (11-15.5)^2/15.5 + (17-12.5)^2/12.5 = 7.2$
- Chi-squared P-value is 0.007 (see next page), so reject H_0 .
- This study provides evidence at the $\alpha = 0.01$ significance level that the probability of improving is greater if the drug is taken than if the placebo is taken.
- Degrees of freedom: $df = 1$ for the 2×2 table.
 - In general it is $(\text{\#rows}-1)(\text{\#columns}-1)$
 - X_s^2 has a χ_1^2 distribution under H_0 .

Chi-square test in R

```
drugs<-read.csv( 'DataSets2/drugs.csv' )  
table(drugs$Treatment, drugs$Response)
```

```
##  
##           Improve NoImprove  
## Drug           15         11  
## Placebo         4         17
```

```
chisq.test(table(drugs$Treatment, drugs$Response), correct=FALSE)
```

```
##  
## Pearson's Chi-squared test  
##  
## data:  table(drugs$Treatment, drugs$Response)  
## X-squared = 7.2037, df = 1, p-value = 0.007275
```

Chi-square test in R

- Note that we used `correct=FALSE` in the `chisq.test` command: this turns off R's default continuity correction, which is used to improve the χ^2 approximation in 2×2 tables.

```
chisq.test(table(drugs$Treatment, drugs$Response))
```

```
##
```

```
## Pearson's Chi-squared test with Yates' continuity correction
```

```
##
```

```
## data:  table(drugs$Treatment, drugs$Response)
```

```
## X-squared = 5.6885, df = 1, p-value = 0.01708
```

What does rejecting H_0 mean?

- Sometimes have to be careful about conclusions in this test. If you reject H_0 with a chi-squared test then that indicates the two variables are **associated**, i.e., not independent; that does not always imply a causal relationship.
- This study provides evidence that patients who take the drug are more likely to improve than patients who take the placebo.
- Here we **controlled** the drug vs. placebo and observed the improvement, so it was reasonable to infer causation. If we had performed a test based on observational data, we could only say that the two variables were associated.

Analysis of Variance (ANOVA)

- How to analyze data when there are k treatments?
 - Same basic assumptions/limitations as t-test:
 - Each population has a normal distribution
 - Samples are independent and random
 - Want to test hypotheses about the population means μ_i
- SD, σ , of each population is the same (or similar) so we can use a pooled SE.
- Goal: test hypotheses such as
 - $H_0: \mu_1 = \mu_2 = \mu_3 = \dots = \mu_k$
 - H_A : at least one of the μ_i 's is different than the others.

ANOVA - The Global F Test

- $H_0: \mu_1 = \mu_2 = \mu_3 = \dots = \mu_k$
- H_A : at least one of the μ_i 's is different than the others.
- To test this hypothesis, we use the test statistic F_s , which is the ratio of the "between-group variation" to the "within-group variation".
- Reject H_0 if observed F_s is "large" (calculate P-value).
- Note: H_A is non-directional for F-Test, but rejection region is one-tailed.

Example

- A random sample of 15 healthy young men are split randomly into 3 groups of 5.
- They receive 0, 20, and 40 mg of the drug Paxil per day for a week, respectively.
- Then their serotonin levels are measured to determine whether Paxil affects serotonin levels.

Hypotheses

- Definitions
 - μ_1 is the mean serotonin level for men on 0 mg of Paxil.
 - μ_2 is the mean serotonin level for men on 20 mg of Paxil.
 - μ_3 is the mean serotonin level for men on 40 mg of Paxil.
- Hypotheses
 - $H_0: \mu_1 = \mu_2 = \mu_3$; mean serotonin levels are the same at all 3 dosage levels (i.e., mean serotonin levels are unaffected by Paxil dose).
 - H_A : The mean serotonin levels of the three groups are not all equal (i.e., serotonin levels are affected by Paxil dose).

One-way ANOVA in R

```
paxil<-read.csv('DataSets2/paxil.csv')
View(paxil)
summary(aov(Serotonin ~ as.factor(Dose), data=paxil))
```

```
##              Df Sum Sq Mean Sq F value    Pr(>F)
## as.factor(Dose)  2   841.9    420.9     8.358 0.00532 **
## Residuals      12   604.3     50.4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- $p = 0.00532$, so reject H_0 .
- This study provides evidence that Paxil intake affects serotonin levels in young men.

How/Why does this work?

- Remember for t-test,
 - take the difference in means: $\bar{y}_1 - \bar{y}_2$
 - divide by the amount of sampling variation it has: $SE_{\bar{y}_1 - \bar{y}_2}$
 - if $\bar{y}_1 - \bar{y}_2$ is big compared to the amount of variability expected by chance, t_s is large; so reject H_0 .
- For F-test,
 - take average squared difference between the group means
 - divide it by the amount of sampling variation it has
 - if the between-group variation is large compared to the amount of variability expected by chance, F is large; so reject H_0

Why not just do t-tests on each pair?

- You could, but the probability of making at least one Type I error is bigger than just the α you use for each test because there can be many possible pairs to test.
- For example when we test $H_0: \mu_1 = \mu_2 = \mu_3$, we are actually testing:
 - $H_{01}: \mu_1 = \mu_2$
 - $H_{02}: \mu_1 = \mu_3$
 - $H_{03}: \mu_2 = \mu_3$
- And we would reject H_0 if any one of H_{01} , H_{02} or H_{03} would be rejected by a t-test with prob of Type I error = 0.05 on first test, 0.05 on second test, etc.

Why not just do t-tests on each pair?

- The more tests you do, the more likely you are to make at least one mistake. (With $k > 7$ you would expect to make at least one mistake).
- Plus, the tests are not independent so it's not even easy to decide what the true Type I error probability is.
- Also, ANOVA lets you get at more complicated patterns.
- So, if you don't reject H_0 , stop. The data provide no evidence in favor of H_A .

Post-hoc testing

- You can perform additional tests to determine which means are different:
- Use the TukeyHSD function in R:

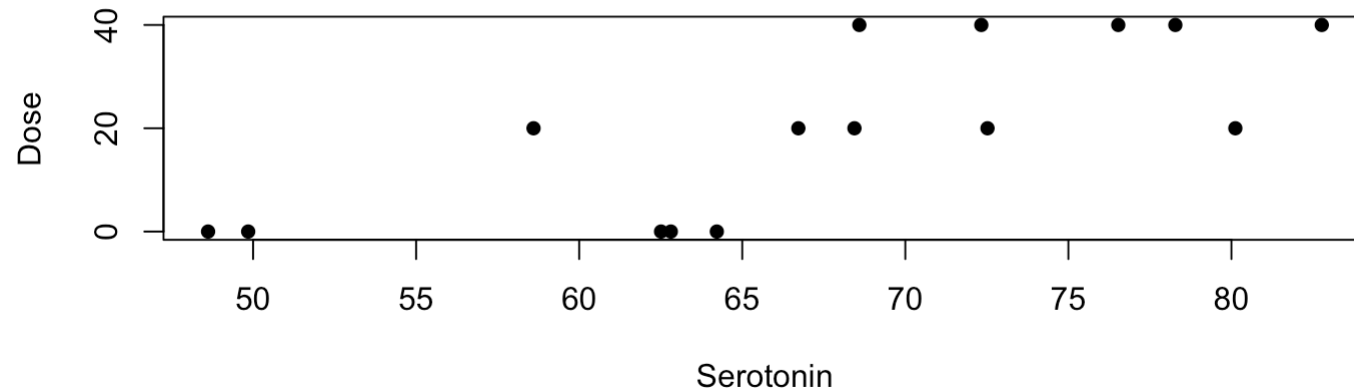
```
TukeyHSD(aov(Serotonin ~ as.factor(Dose), data=paxil))
```

```
## Tukey multiple comparisons of means
## 95% family-wise confidence level
##
## Fit: aov(formula = Serotonin ~ as.factor(Dose), data = paxil)
##
## $`as.factor(Dose)`
##      diff      lwr      upr    p adj
## 20-0  11.678 -0.2961563 23.65216 0.0561323
## 40-0  18.098  6.1238437 30.07216 0.0043776
## 40-20  6.420 -5.5541563 18.39416 0.3571044
```


Post-hoc testing

- There is a significant difference in mean serotonin levels between men receiving a 0mg dose of Paxil, and those receiving a 40mg dose.

```
stripchart(Serotonin ~ as.factor(Dose), data=paxil, pch=16, ylab="Dose")
```



Relationships between quantitative variables

- Often interested in the relationship between two quantitative variables.
- Need methods for:
 - data display
 - describing the relationship

Relationship between variables

- As in the contingency table data, this may be
 - one random sample; observe two quantities
 - one of the variables may be controlled (e.g. dose)
- Examples
 - X = concentration, Y = rate of reaction
 - X = dose, Y = response
 - X = weight, Y = height
 - X = root extension, Y = carbon allocation
 - X = midterm 1 score, Y = midterm 2 score
- If the experimenter controls one of the variables, that is usually labeled X , and the “response” is Y .

Heroin again

- For each cadaver blood taken from heart and from peripheral organs.
- Questions of interest:
 - What is the relationship between the two concentrations?
 - Can we use the heart concentration to predict the peripheral concentration?

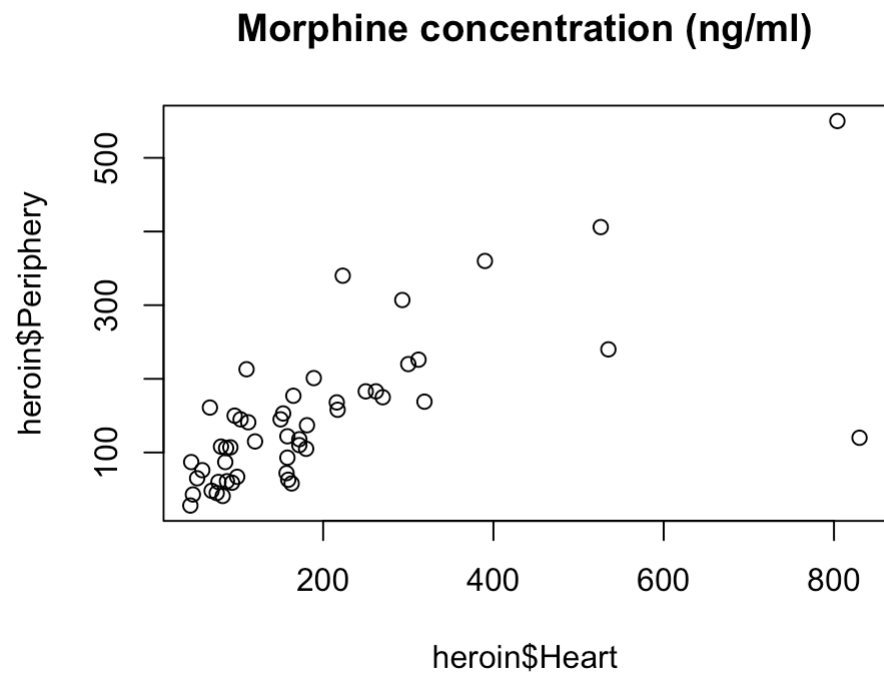
Load heroin data

```
heroin <- read.csv('DataSets2/heroin-regression.csv')  
head(heroin)  
View(heroin)
```

```
##      Heart Periphery  
## 1      67      161  
## 2     110     213  
## 3      45       87  
## 4      96     150  
## 5      80     108  
## 6     103     145
```

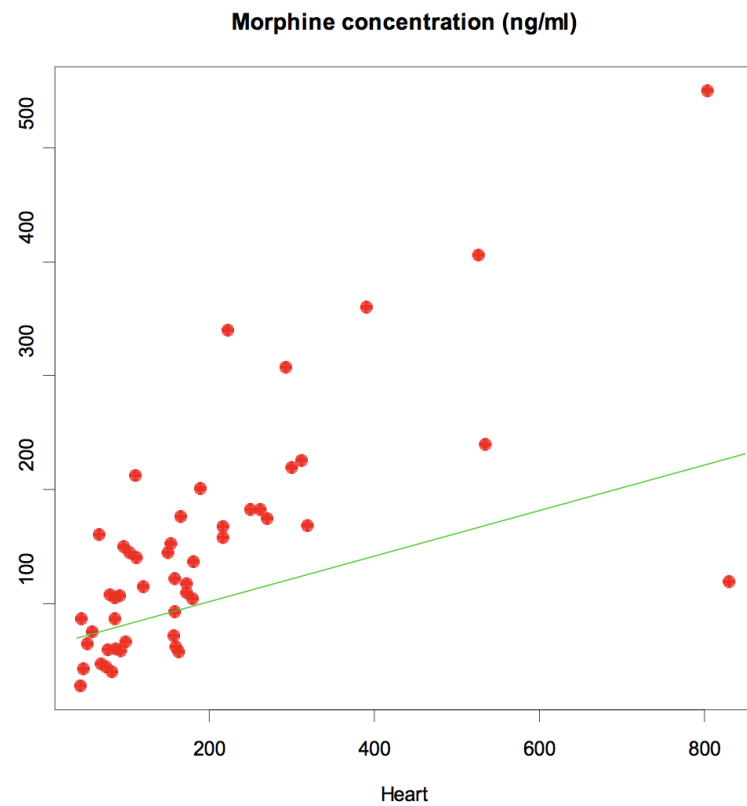
Scatterplot

```
plot(heroin$Heart, heroin$Periphery,  
     main="Morphine concentration (ng/ml)")
```



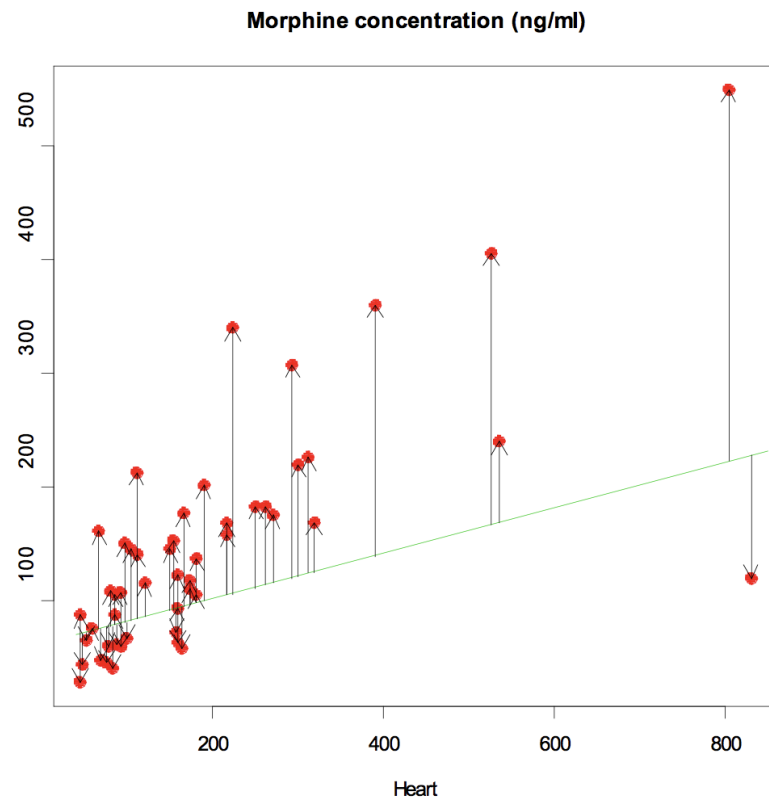
“Line of best fit”

Fit any line: how good is it?



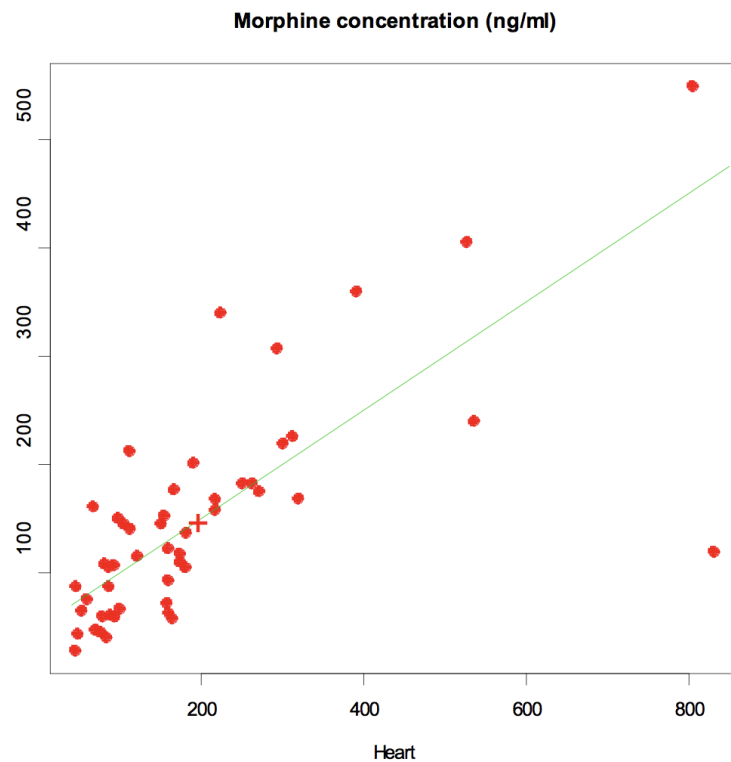
“Line of best fit”

Calculate the distance from each point to the line, square them, and add them all up.



“Line of best fit”

- *Peripheral conc. = 65 + 0.43 Heart conc.*



- Line of best fit minimizes these “sums of squared errors”.

Linear Regression

- A way to look at the relationship between continuous variables
- Back to quantitative data with assumed normal distributions
- Conceptually related to both ANOVA and contingency tables
- “Finding a straight line” through the data
- Data: pairs of observations (X, Y) , so observe:
 - $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$

What does “best” mean?

- For each observed value of x there is a value of y predicted by the straight line, which we write as

$$\hat{y} = b_0 + b_1x$$

- This is usually not exactly the same as the observed y for that x .
- Compare the actual vs. predicted y ; difference is called residual:
- Residual = $y - \hat{y}$ (one for each pair of observations, i.e., n residuals)

Regression in R

```
summary(lm(Heart ~ Periphery, data=heroin))
```

```
##
```

```
## Call:
```

```
## lm(formula = Heart ~ Periphery, data = heroin)
```

```
##
```

```
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
```

```
## -196.95  -54.70   -5.40   25.11  668.61
```

```
##
```

```
## Coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)  20.3604    30.3705   0.670   0.506
```

```
## Periphery    1.1753     0.1699   6.916 9.9e-09 ***
```

```
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
```

```
## Residual standard error: 121.4 on 48 dearees of freedom
```

Regression output

- Coefficients are the intercept and slope of the regression line.
- Significance of the slope means that there is a linear relationship between the variables.

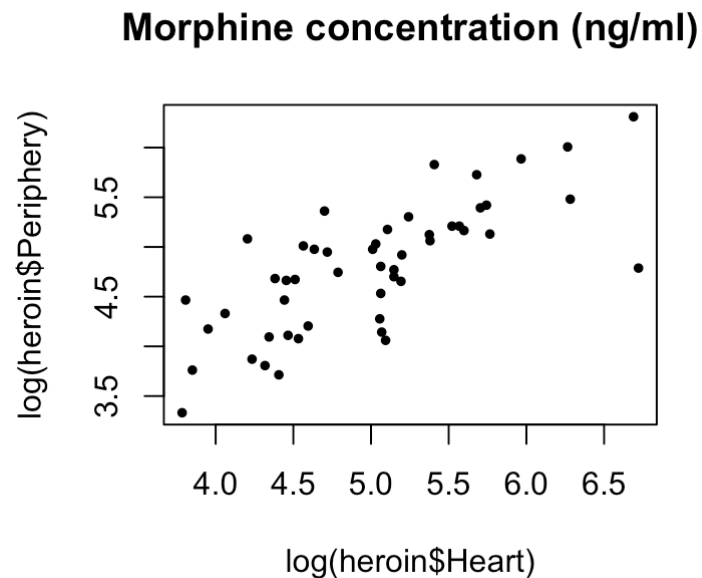
```
summary(lm(Heart ~ Periphery, data=heroin))$coefficients
```

| ## | | Estimate | Std. Error | t value | Pr(> t) |
|----|-------------|-----------|------------|-----------|--------------|
| ## | (Intercept) | 20.360409 | 30.3705099 | 0.6704006 | 5.058153e-01 |
| ## | Periphery | 1.175255 | 0.1699363 | 6.9158543 | 9.896584e-09 |

Aside - log transformation

- The relationship between the Heart and Periphery variables is more linear when each variable is log transformed.

```
plot(log(heroin$Heart), log(heroin$Periphery), pch=20, cex=0.8,  
     main="Morphine concentration (ng/ml)")
```



Correlation Coefficient

- The correlation coefficient, ρ , measures the strength of linear association between two normally distributed variables.
- $r = \hat{\rho}$ has the same sign as the slope of the regression model.
- It is always between -1 and +1 and r has no units.
- It will be 1 only if the data points are exactly on a straight line so that all the residuals are 0.
- It will be 0 when there is no linear relationship between X and Y .
- A value “close” to 1 indicates a “tight” linear relationship between X and Y .

Correlation

- If r is large in magnitude (close to 1) we say X and Y are highly “correlated”.
 - “Negatively correlated” when r is negative,
 - “Positively correlated” when r is positive.
- If $r = 0$ there is no linear relationship; we say X and Y are “uncorrelated” (Note: uncorrelated is not the same thing as independent!)
- The word “uncorrelated” is also used to describe data with small but nonzero values of r .

Testing $H_0: \rho = 0$

- Test H_0 the same way (questions, distributions, conclusions) as above:
 - Is there a linear relationship between Y and X?
- Let ρ be the correlation coefficient between X and Y for the population.
 - $H_0: \rho = 0$ There is no linear relationship between Y and X.
 - $H_A: \rho \neq 0$ There is a linear relationship between Y and X. (or H_A could be directional, e.g., $H_A: \rho > 0$).

Correlation in R

- Default is “Pearson”: for normally distributed variables

```
cor(heroin$Heart, heroin$Periphery)
```

```
## [1] 0.7064758
```

- “Spearman” is a rank-based approach (doesn’t require normality).

```
cor(heroin$Heart, heroin$Periphery, method='spearman')
```

```
## [1] 0.7625885
```

Test for significance

```
cor.test(heroin$Heart, heroin$Periphery)

##
## Pearson's product-moment correlation
##
## data: heroin$Heart and heroin$Periphery
## t = 6.9159, df = 48, p-value = 9.897e-09
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
## 0.5329257 0.8229865
## sample estimates:
## cor
## 0.7064758
```

Assignment: R Markdown

- RStudio provides a framework for producing documents within R: R Markdown
- R Markdown uses a simple syntax to produce documents containing:
 - comments
 - code
 - output
- By default, running an R Markdown (.Rmd) file will produce HTML output, although other output formats can be used.
- See the website below for more details, and click the “Get Started” link to get a video overview.

<http://rmarkdown.rstudio.com/>

Creating an R Markdown document

- In RStudio, select the 'New File' then 'R Markdown...' from the 'File' menu.
- Choose 'Document' and give it a name (leave the output format as "HTML"), then click 'OK'.
- This creates a basic R Markdown document containing some sample code:
 - the top of the file contains some basic information
 - blocks of R code have `` `` {r}` in front, and `` ``` after: these are run when you click "Knit" (see next slide), and the output appears in the HTML file
 - everything else in the file is treated as text to be displayed

Creating an R Markdown document

- Various text effects can be achieved using markdown syntax (e.g., “##” for headings): see the link below for (many) more options.
- Clicking the “Knit” button in RStudio (above your code window) will run the code, and open a built-in browser that displays the processed document as rendered HTML.
- This file is saved in a folder within your working directory, and can be opened in any web browser, exported to other formats etc.
- A big advantage of R Markdown is *reproducibility* - anyone with the .Rmd file and associated data can *exactly* reproduce the results in the document.

http://rmarkdown.rstudio.com/authoring_basics.html