Outline

1. Corrective actions and Nonparametric methods
   - Data transformations
   - Mann-Whitney test
Assessing assumptions

The t-test assuming equal variances is very sensitive to dependence, moderately robust against unequal variance if \( n_1 \approx n_2 \), but much less robust if \( n_1 \) and \( n_2 \) are quite different (e.g. differ by a ratio of 3 or more).

robust against nonnormality.

Corrective actions for 2 independent samples:
Fundamental changes if problem with independence (...
Welch t-test if \( \sigma_1 \) and \( \sigma_2 \) differ by 3-fold or more or \( n_1 \) and \( n_2 \) differ by 3-fold or more.
If non-normal distributions:

1. try a data transformation,
2. or switch to a non-parametric test: Mann-Whitney test.
A veterinarian wishes to know if the presence of a certain fetlock disorder in race horses affects their selling price at auction. Data on 8 horses that have the disorder, and 11 that do not (in $)

**With Disorder:** 5000, 6000, 14100, 49000, 7000, 26000, 2000, 2200

**Without Disorder:** 27000, 14000, 11500, 19000, 9500, 40000, 75000, 9000, 14500, 50000, 30500
Both samples are skewed right. Look at the log-values of the prices:

```r
> dis
[1]  5000  6000 14100  49000  7000 26000  2000  2200
> log(dis)
[1]  8.52  8.70  9.55  10.80  8.85 10.17  7.60  7.70

> nod
[1] 27000 14000 11500 19000  9500 40000  75000  9000 14500 50000 30500
> log(nod)
```

Could we do the t-test on log-values instead?

If the price tends to go down with the fetlock disorder, then the log(price) also tends to be lower with the disorder than without (and vice versa).
Theoretical Quantiles
Sample Quantiles
log of price,
without disorder

log of price,
with disorder

Frequency
7 8 9 10 11
0.0 1.0 2.0 3.0
log of price

log of price,
with disorder

log of price,
without disorder

Frequency
7 8 9 10 11
0.0 1.0 2.0 3.0
log of price

Theoretical Quantiles
Sample Quantiles
log of price,
without disorder

log of price,
with disorder
T-test on the log-transformed prices

The distribution of log-prices looks beautifully normal for both samples! Welch t-test on the log-transformed prices:

dis = c(5000, 6000, 14100, 49000, 7000, 26000, 2000, 2200)
nod = c(27000, 14000, 11500, 19000, 9500, 40000, 75000, 9000, 14500, 50000, 30500)

> t.test(log(dis), log(nod))

Welch Two Sample t-test

data: log(dis) and log(nod)
t = -2.1955, df = 10.951, p-value = 0.05059
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  -1.988749799 0.003048454
sample estimates:
mean of x mean of y
  8.985856  9.978706

Conclusion: moderate evidence that the auction prices tend to be lower with the fetlock disorder than without (p=0.051).
Transformations

Log transformation:
helps when the distributions are skewed right,
only when all values are positive

Square-root transformation:
helps when distributions are moderately skewed right,
only when all values are $\geq 0$ (zeros are okay)

Apply the same transformation (here: take the log) to all values in both samples.

Choose the transformation in order to satisfy assumptions, not based on the resulting p-value.

Confidence intervals on the original scale (not log, not transformed) are more difficult to get.
What if...

The data are too skewed and no transformation can help?

For instance: a transformation might help make one sample look normally distributed but make the other sample look worse.

Third option: use a ‘non parametric’ test, here test that does not assume the normal distribution: the Mann-Whitney test.
Mann-Whitney test (aka Wilcoxon rank sum test)

Analogous to the Wilcoxon signed-rank test (for paired samples) but here for two independent samples.

No distribution assumption, but still assume independence.

Main idea: look at the ranks of the observations.

Example: Does soil respiration affect plant growth? Soil cores taken from 2 locations in a forest: under an opening in the forest canopy (“gap”) and at a nearby area under heavy tree growth (“growth”). Measured: amount of carbon dioxide given off by each soil core (mol CO$_2$/g soil/hr). Data:

<table>
<thead>
<tr>
<th>Gap</th>
<th>22</th>
<th>29</th>
<th>13</th>
<th>16</th>
<th>15</th>
<th>18</th>
<th>14</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Growth</td>
<td>17</td>
<td>20</td>
<td>170</td>
<td>315</td>
<td>22</td>
<td>190</td>
<td>64</td>
<td></td>
</tr>
</tbody>
</table>
Gap data: distribution has normal shape,
Growth data: skewed right.
Welch t-test not recommended, but there is another way!
pdf("lec15-01.pdf",width=5,height=5)
gap  =c(22,29,13,16,15,18,14,6)
growth=c(17,20,170,315,22,190,64)
layout(matrix(1:4,2,2))
par(mar=c(3.1,3.1,1.5,.5), mgp=c(1.8,.4,0), tck=-0.01, las=1,bty="n")
hist(gap, xlim=c(5,400))
hist(growth, xlim=c(5,400), breaks=10)
qqnorm(gap ,pch=16)
qqnorm(growth,pch=16)
dev.off()
Mann-Whitney test

$H_0$: the 2 populations have the same distribution. Soil respiration has the same distribution in the 2 locations, with $\mu_1 = \mu_2$ in particular.

$H_A$: soil respiration does not have the same distribution in the 2 populations. Test most sensitive to a shift between the 2 distributions, so it’s usually assumed that $H_A$ is: ‘the 2 distribution have different means’.
Mann-Whitney test

Rank the observations, calculate:

\[ U_1 = \# \text{ of observations in group 2 that are smaller} \]

\[ U_2 = \# \text{ of observations in group 1 that are smaller} \]

and summarize the data by \( U = \max\{U_1, U_2\} \).

If \( H_0 \) is true, then \( U \) has a Wilcoxon distribution (does not depend on the common distribution of the data).
soil respiration rank

|   |   |   |   |   |   |   |   |   |   |   |   | 6 | 14 |
|---|---|---|---|---|---|---|---|---|---|---|---|    |    |
| gap | △ | △ | △ | △ | △ | △ | △ | △ | △ | △ |   |   |    |
| growth |   |   |   |   |   |   |   |   |   |   |   |   |   |    |


\[
U_1 \text{ (gap)} = 0 + 0 + 0 + 0 + 0 + 1 + 2.5 + 3 = 6.5
\]

\[
U_2 \text{ (growth)} = 5 + 6 + 6.5 + 8 + 8 + 8 + 8 = 49.5 \text{ so } U = 49.5.
\]

(for ties: count 0.5)

To double-check: \(U_1 + U_2 = n_1 \times n_2\) always.

Here yes: \(49.5 + 6.5 = 7 \times 8 = 56\)

If \(H_0\) is true: assignment of ranks to sample is completely random, so expectation: \(U_1\) and \(U_2\) should be similar, i.e. both intermediate, i.e. both about \(n_1 \times n_2 / 2\) (= 28 here).

\(U = \max\{U_1, U_2\}\) expected to be moderate.

More extreme in the direction of \(H_A\): imbalance between \(U_1\) and \(U_2\) (one small, one large), i.e. large \(U\).
Mann-Whitney test

3. We got $U = 49.5$, more extreme = larger, so
p-value=$\mathbb{P}\{U \geq 49.5\}$.

Table E, $n_1 = 8$ and $n_2 = 7$: critical (minimum) $U$ is 46 for rejecting at $\alpha = 0.05$, and 50 at $\alpha = 0.01$

So here $0.01 < p\text{-value} < 0.05$

4. Conclusion: we have moderate evidence that the soil respiration distribution differs between the two locations.

Soil respiration has a higher mean in the area under heavy tree growth, than in the area under the opening of the forest canopy.

Note: Table E has no number listed for $n_1 = 3$ and $n_2 = 4$: we can never reject $H_0$ at $\alpha = 0.05$. 
One-sided Mann-Whitney test

\( H_A \): distribution shift with \( \mu_1 > \mu_2 \) for instance.

First check that the data go in the same direction as \( H_A \), i.e. check that \( U_1 > U_2 \) if testing \( H_A: \mu_1 > \mu_2 \).

- If not: p-value > 0.50.
- If so: p-value is half as much as what it would be for a two-sided test.
wilcox.test() in R

> gap
[1] 22 29 13 16 15 18 14  6
> growth
[1] 17 20 170 315 22 190  64

> wilcox.test(gap, growth)

Wilcoxon rank sum test with continuity correction

data:  gap and growth
W = 6.5, p-value = 0.015
alternative hypothesis: true location shift is not equal to 0

Warning message:
In wilcox.test.default(gap, growth) :
cannot compute exact p-value with ties
Warnings and Assumptions

If there are ties, the table gives approximation only.

The test does not work well if the variances are very different.

To interpret $H_A$ as simply $\mu_1 \neq \mu_2$ rather than ‘the 2 distributions are different’, we actually need to assume that when the 2 distributions differ, they only differ by their means — not variances.

The Mann-Whitney test is less powerful than the t-test (when both are applicable) for small sample sizes, but almost as powerful for large sample sizes.

Try transformation + t-test first: more powerful if applicable. Otherwise use Mann-Whitney.