The vent damper example[1]

STA442/2101 Fall 2014

[1] See last slide for copyright information.

The Vent damper data

Based on a Minitab data set, but all the numbers are different

When a furnace is off, the chimney does not need to be open. Maybe, closing it can conserve energy. A vent damper is a kind of barrier that closes the chimney. Dampers on fireplaces are manual, but in modern heating systems they are automatic. Automatic vent dampers can be either electrical or thermal. Which kind saves more energy?

Vent damper continued

Forty houses were randomly assigned to have electrical vent dampers installed, and 50 were randomly assigned to have thermal vent dampers. Average daily energy consumption was measured during two consecutive weeks in the winter. For one week the vent damper was active (turned on), and the other week it was inactive (turned off). Within damper type, equal numbers of houses were randomly assigned to have the damper on during the first week and the second week.

The cases are houses. The three variables in the file are energy consumption with vent damper active, energy consumption with vent damper inactive, and type of damper. The first question is whether it's true that energy consumption is less when the damper is active. For this analysis, we will ignore whether the vent damper is electrical or thermal.

The main question (for now)

Does the vent damper affect energy consumption, and if so, by how much?

- Let's do a t-test.
- It's natural to calculate a *difference* in energy consumption for each house, and test whether the mean difference equals zero.
- Or, we could do a two-sample t-test with $n_1 = n_2$.
- Which one is better?
- A test implies a model; compare the models.

Model for the matched t-test

Independently for $i = 1, \ldots, n$, observe (X_i, Y_i).

- $X_i \sim N(\mu_1, \sigma_1^2)$, $Y_i \sim N(\mu_2, \sigma_2^2)$.
- $Cov(X_i, Y_i) = \sigma_{12}$.
- Calculate Differences $D_i = X_i - Y_i$
- Matched t-test on $D_1, \ldots, D_n$
- $H_0 : \mu = 0$, where $\mu = E(D_i) = \mu_1 - \mu_2$

Test statistic is

$$T_1 = \frac{\sqrt{n}(\overline{D} - 0)}{S}$$

with $df = n - 1$.

Independent t-test

Correct if $\sigma_1^2 = \sigma_2^2$ and $\sigma_{12} = 0$

$$T_2 = \frac{\overline{X} - \overline{Y}}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

with $df = n_1 + n_2 - 2$, where

$$S_p^2 = \frac{\sum_{i=1}^{n_1} (X_i - \overline{X})^2 + \sum_{i=1}^{n_2} (Y_i - \overline{Y})^2}{n_1 + n_2 - 2}$$

Comparing the Tests

$$T_1 = \frac{\sqrt{n}(\overline{D} - 0)}{S}, \quad df = n - 1$$

$$T_2 = \frac{\overline{X} - \overline{Y}}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \quad df = 2(n - 1)$$

- The two-sample test pretends it has twice the degrees of freedom.
- Could cause worry about inflated Type I error rate
- But both critical values go to $z_{\alpha/2}$ as $n \to \infty$.
- For example, for $n = 100$, $t_{0.975}(99) = 1.98$ while $t_{0.975}(198) = 1.97$.
- So if there is a problem with df, it will be for small samples.

Comparing the Test Statistics

$$T_1 = \frac{(\overline{D} - 0)}{S/\sqrt{n}}, \quad df = n - 1$$

$$T_2 = \frac{\overline{X} - \overline{Y}}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \quad df = 2(n-1)$$

- $\overline{D} = \frac{1}{n}\sum_{i=1}^{n}(X_i - Y_i) = \overline{X} - \overline{Y}$
- So the numerators are the same.
- Compare denominators

One-Sample (Matched) t-Test

$$
\begin{aligned}
S^2/n &= \frac{1}{n(n-1)} \sum_{i=1}^{n} (D_i - \overline{D})^2 \\
&= \frac{1}{n(n-1)} \sum_{i=1}^{n} \left(X_i - Y_i - (\overline{X} - \overline{Y})\right)^2 \\
&= \frac{1}{n(n-1)} \sum_{i=1}^{n} \left((X_i - \overline{X}) - (Y_i - \overline{Y})\right)^2 \\
&= \frac{1}{n} \left[\frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n-1} - 2 \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{n-1} \right. \\
&\quad \left. + \frac{\sum_{i=1}^{n} (Y_i - \overline{Y})^2}{n-1} \right] \\
&= \frac{1}{n} \left[S_x^2 - 2S_{xy} + S_y^2 \right]
\end{aligned}
$$

where S_{xy} is the sample covariance.

Two-Sample (Independent) t-Test

With $n_1 = n_2 = n$

$$
\begin{aligned}
S_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right) &= \frac{(n_1 - 1)S_x^2 + (n_2 - 1)S_y^2}{n_1 + n_2 - 2} \left(\frac{1}{n_1} + \frac{1}{n_2} \right) \\
&= \frac{(n-1)(S_x^2 + S_y^2)}{n + n - 2} \left(\frac{2}{n} \right) \\
&= \frac{(n-1)(S_x^2 + S_y^2)}{2(n-1)} \left(\frac{2}{n} \right) \\
&= \frac{S_x^2 + S_y^2}{n}
\end{aligned}
$$

Comparing (Squared) Denominators

$$S_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right) = \frac{1}{n}\left[S_x^2+S_y^2\right]$$

$$S^2/n = \frac{1}{n}\left[S_x^2-2S_{xy}+S_y^2\right]$$

- If covariance is zero, they are the same.
- If covariance is negative
  - Denominator of two-sample t is too small.
  - Value of t too large.
  - Null hypothesis rejected too often.
- If covariance is positive
  - Denominator of two-sample t is too large.
  - Value of t too small.
  - Null hypothesis *less* likely to be rejected.
  - If H_0 is false, expect loss of power.

Covariance should be positive: Why?

A more detailed model

Independently for $i = 1, \ldots, n$, observe (X_i, Y_i) where

$$\begin{aligned} X_i &= \delta + Z_i + \epsilon_{i1} \\ Y_i &= Z_i + \epsilon_{i2} \end{aligned}$$

- X_i is the measurement with vent damper active.
- δ is the effect of having the vent damper active.
- Z_i reflects characteristics of the individual house (surface area, insulation, habits of the occupants, etc.). $Z_i \sim N(\mu_z, \sigma_z^2)$.
- ϵ_{ij} reflects other influences not specific to the house (measurement error, weather etc.). $\epsilon_{ij} \sim N(0, \sigma_j)$ for $j = 1, 2$.
- Z_i and ϵ_{ij} all independent.

Covariance is positive

And could be quite large

$$\begin{aligned} X_i &= \delta + Z_i + \epsilon_{i1} \\ Y_i &= Z_i + \epsilon_{i2} \end{aligned}$$

$$Cov(X_i, Y_i) = Var(Z_i)$$

- Notice Z_i cancels in $D_i = X_i - Y_i$.
- A lot of extraneous variance is removed.
- Each house serves as its own control.
- And $\delta = \mu_1 - \mu_2$.

Read the data

```
> furnace = read.table("http://www.utstat.utoronto.ca/~brunner/
            appliedf14/code_n_data/lecture/damper.data")
            # URL is all on one line
> head(furnace)

  damper active inactive
1    EVD  10.71    11.46
2    EVD   7.45     7.91
3    EVD   9.25    10.48
4    EVD  15.91    16.97
5    EVD   7.47     7.88
6    EVD   7.29     7.75
```

Type `furnace` to look at the whole data set.

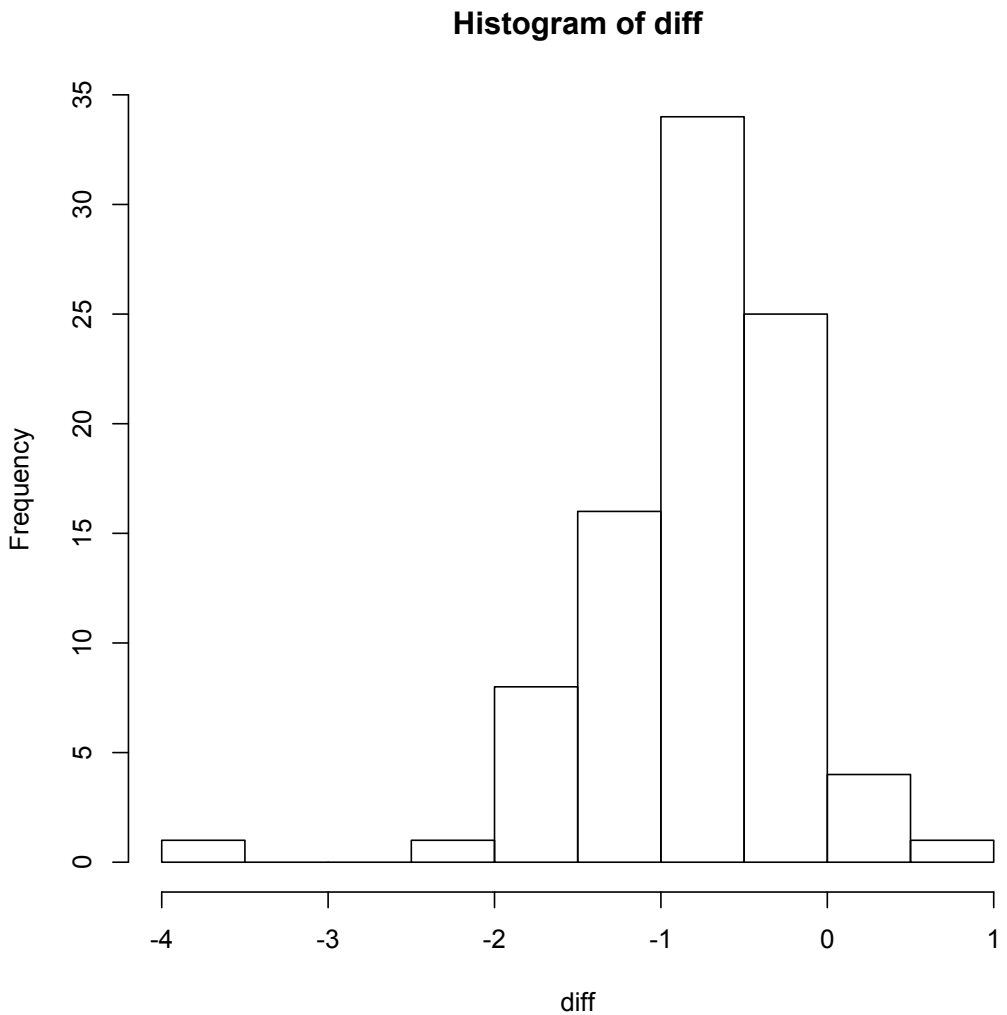
Summary of Difference score

```
> attach(furnace) # Make variable names available
> diff = active-inactive; summary(diff)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-3.9800 -1.0550 -0.7100 -0.7747 -0.4175  0.8700
```

Histogram of Difference score

```
> hist(diff)
```



Check the number

```
> sort(diff) # Looks like just one possible outlier

[1] -3.98 -2.29 -1.95 -1.94 -1.59 -1.57 -1.54 -1.53 -1.53 -1.52 -1.38 -1.34 -1.
[18] -1.11 -1.11 -1.10 -1.09 -1.09 -1.06 -1.04 -1.03 -1.01 -0.99 -0.98 -0.97 -0
[35] -0.83 -0.81 -0.80 -0.80 -0.79 -0.75 -0.75 -0.74 -0.73 -0.72 -0.72 -0.70 -0
[52] -0.67 -0.66 -0.62 -0.61 -0.60 -0.59 -0.58 -0.57 -0.56 -0.49 -0.47 -0.47 -0
[69] -0.41 -0.38 -0.37 -0.37 -0.37 -0.32 -0.30 -0.30 -0.29 -0.27 -0.26 -0.23 -0
[86]  0.17  0.25  0.29  0.38  0.87
```

```
> (1:90)[diff==min(diff)]

[1] 90
```

It's House number 90.

How unusual is that observation?

Standardize

If D_i are normal, $Z_i = \frac{D_i - \overline{D}}{S_d}$ are approximately standard normal

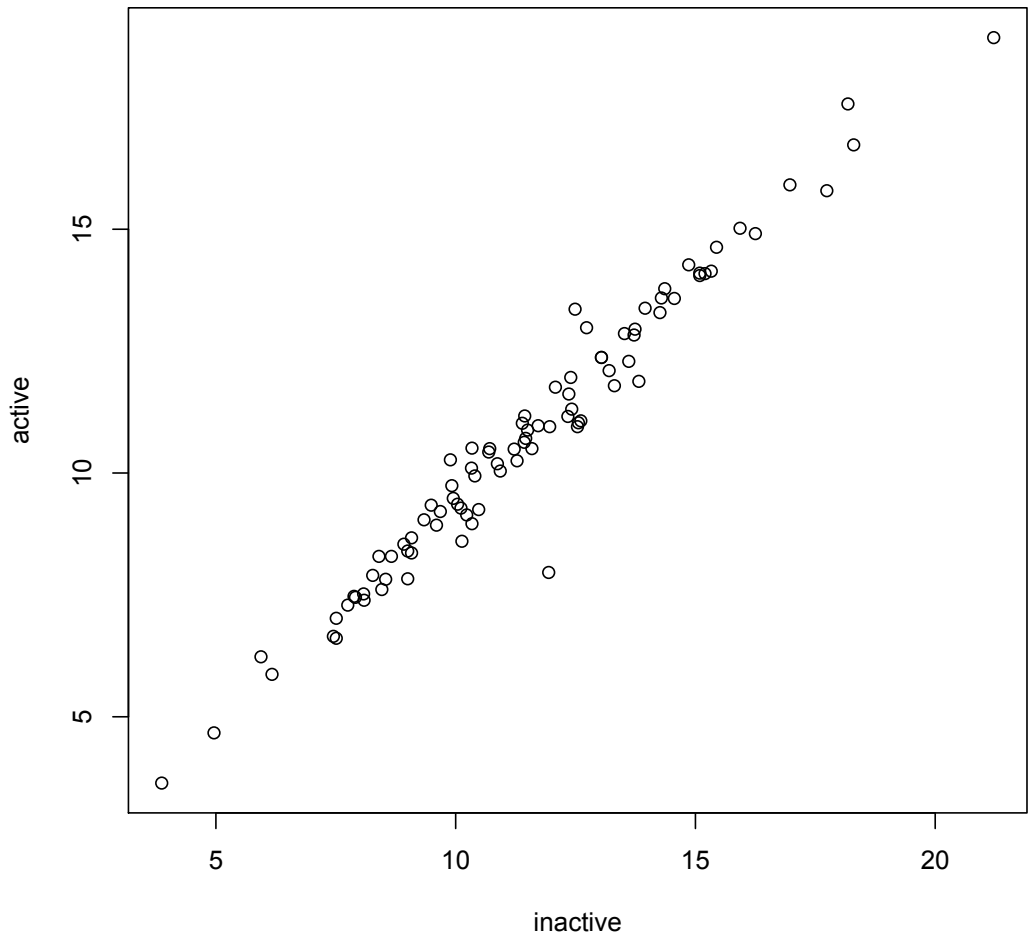
```
> Z = (diff-mean(diff))/sd(diff); sort(Z)[1:5]
```

```
[1] -5.177325 -2.447600 -1.898424 -1.882272 -1.316944
```

$Z = -5.177$ is really unusual. Homework: What's the probability of getting one or more this big in absolute value for $n = 90$ independent standard normal data?

Does it look like an outlier in two dimensions?

```
> plot(inactive,active)
```



What is unusual about House Number 90?

- Looking in the Minitab handbook, find *nothing* remarkable about that house on any of about 10 variables.
- In a real data analysis job, look harder.
- I don't want to throw it out unless I know *why* it's different.
- Do the analyses with and without this house. If the conclusions are similar, we are happy.

Matched t-test with the full data set

```
> t.test(diff) # Matched t

One Sample t-test

data:  diff
t = -11.8705, df = 89, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -0.9043367 -0.6449966
sample estimates:
 mean of x
-0.7746667
```

Conclusion: Use of the vent damper reduces average energy consumption.

Two-sample t-test with the full data set

```
> t.test(active,inactive,var.equal = T) # Two-sample (independent) t

Two Sample t-test

data:  active and inactive
t = -1.7437, df = 178, p-value = 0.08294
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.6513757  0.1020424
sample estimates:
mean of x mean of y
 10.70844  11.48311
```

Conclusion: These results are consistent with no effect of vent damper.

Matched t-test with outlier deleted

Including outlier, had $t = -11.8705$

```
> activeDel   = active[-90] # Could have said active[1:89]
> inactiveDel = inactive[-90] # Can give it a list, like active[-c(1,2,
> # But watch out for missing values!
> diffDel = activeDel-inactiveDel
> t.test(diffDel)

One Sample t-test

data:  diffDel
t = -13.421, df = 88, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -0.8480265 -0.6292769
sample estimates:
 mean of x
-0.7386517
```

Two-sample t-test with outlier deleted

Including outlier, had $t = -1.7437$

```
> t.test(activeDel,inactiveDel,var.equal = T)

Two Sample t-test

data:  activeDel and inactiveDel
t = -1.6481, df = 176, p-value = 0.1011
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.6231323  0.1458289
sample estimates:
mean of x mean of y
 10.73933  11.47798
```

Comments

- As expected, the two-sample t-test was less sensitive.
- Because the covariance between energy consumption with vent damper active and vent damper inactive was large, the two-sample t-test was *much* less sensitive.
- Deleting the outlier actually made the results more convincing, even though it made the largest contribution to the observed difference between means.
- If the confidence interval for the effect of vent damper could be converted to dollars, it would be very meaningful.

What about normality?

The t-test assumes normality,

- If the outlier is dropped, the data might be normal, maybe.
- Try a sign test.
- Then discuss robustness.

Sign test

- Under the null hypothesis of no effect, energy consumption with vent damper active and vent damper inactive are identically distributed.
- Therefore if the distributions are continuous, $Pr\{X_i > Y_i\} = Pr\{X_i < Y_i\} = \frac{1}{2}$.
- (What do you have to assume about the *joint* distribution to actually show this?)
- Use a Bernoulli model and test $H_0 : \theta = 0.5$

Calculate the sign test

```
> # Sign test
> neg = length(diff[diff<0]); pos = length(diff[diff>0])
> neg; pos

[1] 85
[1] 5
```

```
> p = neg/90
> Z = sqrt(90)*(p-1/2)/sqrt(p*(1-p)); Z

[1] 18.40716
```

There can be no doubt of the effect.

Also, the t-test is *robust* with respect to normality

Compare $T = \frac{\sqrt{n}(\overline{D}-0)}{S}$ to $Z_n = \frac{\sqrt{n}(\overline{D}-0)}{S}$.

- Central Limit Theorem says $Z_n \xrightarrow{d} Z \sim N(0,1)$.
- And as $df \to \infty$, the t distribution becomes standard normal.
- So for large enough samples, the assumption of normality used in the derivation of the t-test is not actually necessary.

Moral of the story: It's always nice when the model is realistic, but when the model is unrealistic it may or may not matter. We have to study the consequences of model incorrectness.

Copyright Information