Paired and Two Sample Testing

Sumanta Basu

Logistics

- Reading list updated
- HW6 (Confidence Intervals) posted, due Monday of next week
- HW7 (Hypothesis Tests) will be posted tomorrow, due Saturday next week
- Second exploration (inference in your research field) will be posted this weekend, due in two weeks
- **Prelim2**: Nov 7 (Thu), 7:30-9:30 pm
- **Make-up Prelim2**: Nov 8 (Fri), time and place TBA (let me know by this weekend if you need to take make-up prelim 2)
- Prelim 2 from last year posted on Blackboard
- Next 3 lectures: examples of confidence intervals and hypothesis tests (one sample, two sample, categorical data)

Prelim 1 Scores

```
mean(prelim1)
```

```
## [1] 63.08511
```

```
sd(prelim1)
```

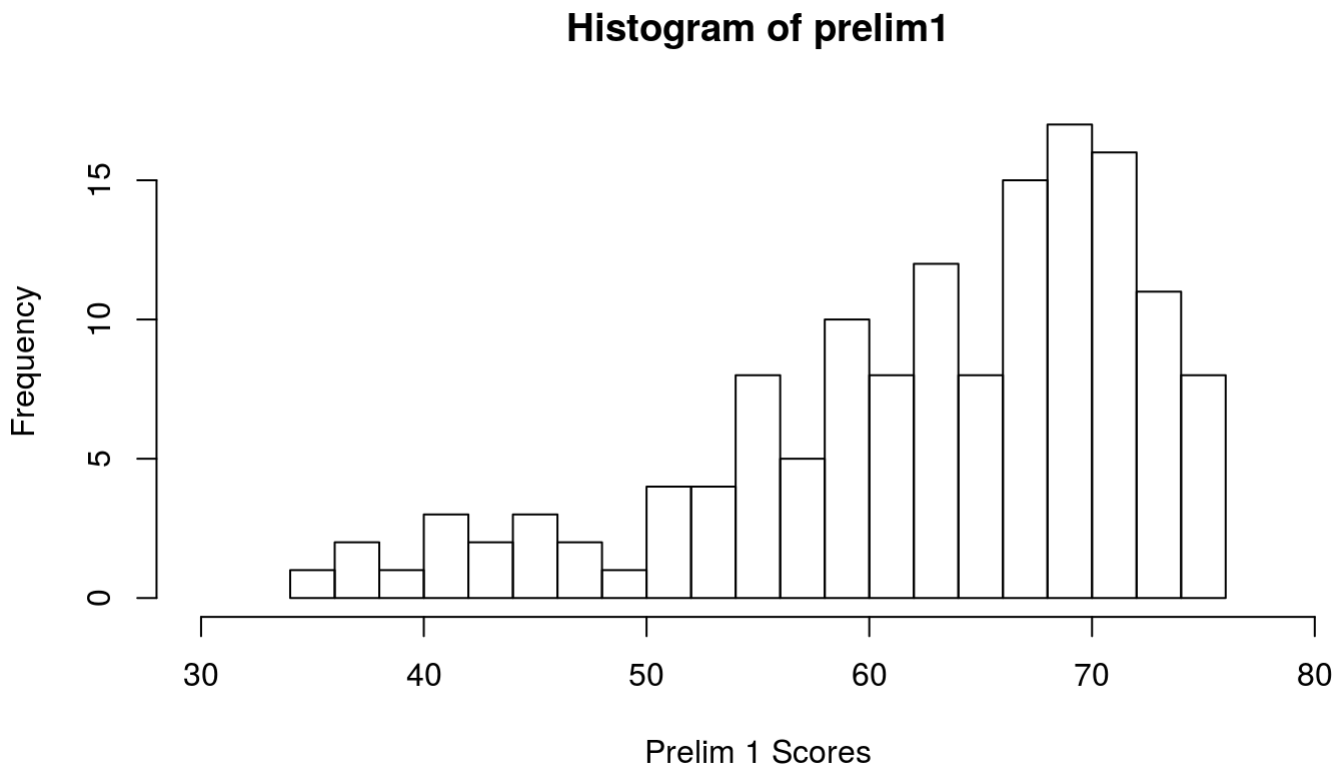
```
## [1] 9.394329
```

```
summary(prelim1)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   34.50   57.50   66.00   63.09   70.00   75.00
```

Prelim 1 Scores

```
hist(prelim1, breaks = 20, xlab = 'Prelim 1 Scores', xlim = c(30, 80))
```



Inference on avg prelim1 score (in %)

```
prelim1.pct = 100*prelim1/75
mean(prelim1.pct)
```

```
## [1] 84.11348
```

```
sd(prelim1.pct)
```

```
## [1] 12.52577
```

```
length(prelim1.pct) # n = sample size
```

```
## [1] 141
```

- Can we give a confidence interval for *population* average prelim1 score of BTRY6010 students?
- Test the hypothesis: population average prelim1 score is 83%

Confidence Interval of Prelim 1 Scores

```
xbar = mean(prelim1.pct); s = sd(prelim1.pct); n = length(prelim1.pct)
alpha = 0.05
std_error = s/sqrt(n)
multiplier = -qnorm(alpha/2) # Qn: Should we choose normal or t?
margin_error = multiplier * std_error
xbar - margin_error  # lower confidence limit
```

```
## [1] 82.04599
```

```
xbar + margin_error # upper confidence limit
```

```
## [1] 86.18096
```

Questions:

- Assumption?
- How to interpret this *interval*?
- How to interpret the *confidence level*?
- **TRUE/FALSE**: There is a 95% chance that μ is between 82.05 and 86.18.

Test if $\mu = 83$

Let μ be the population average prelim 1 score of BTRY 6010 students

$$H_0 : \mu = 83 \qquad vs. \qquad H_A : \mu \neq 83$$

Sample statistic: $\bar{X}$, its realized value = 84.11, null value: $\mu_0 = 83$

Test statistic:

$$Z = \frac{\bar{X} - \mu_0}{SE(\bar{X})}$$

Under null, test statistic Z is distributed as approximately $N(0, 1)$ [**null distribution**]

Realized value of test statistic:

$$\frac{84.11 - 83}{12.53/\sqrt{141}} = 1.0519$$

Test if $\mu = 83$

Calculate p-value [draw picture on board]

```
area.left.tail = pnorm(-1.0519) # P(Z < -1.0519)
area.right.tail = pnorm(1.0519, lower.tail = FALSE) # P(Z > 1.0519)
pval = area.left.tail + area.right.tail # essentially double the area in left tail
pval
```

```
## [1] 0.2928454
```

Since $p \geq \alpha$, we fail to reject H_0.

Conclusion: At 5% level of significance, we do not have sufficient evidence (p-value = 0.29) that the population average prelim 1 score (in %) is different from 83.

Questions: Assumptions? What if we tested $H_A : \mu > 83$? or $H_A : \mu < 83$? What can we do increase power of this test? Is this difference (84.11 vs. 83) *practically* significant?

Comparing two means

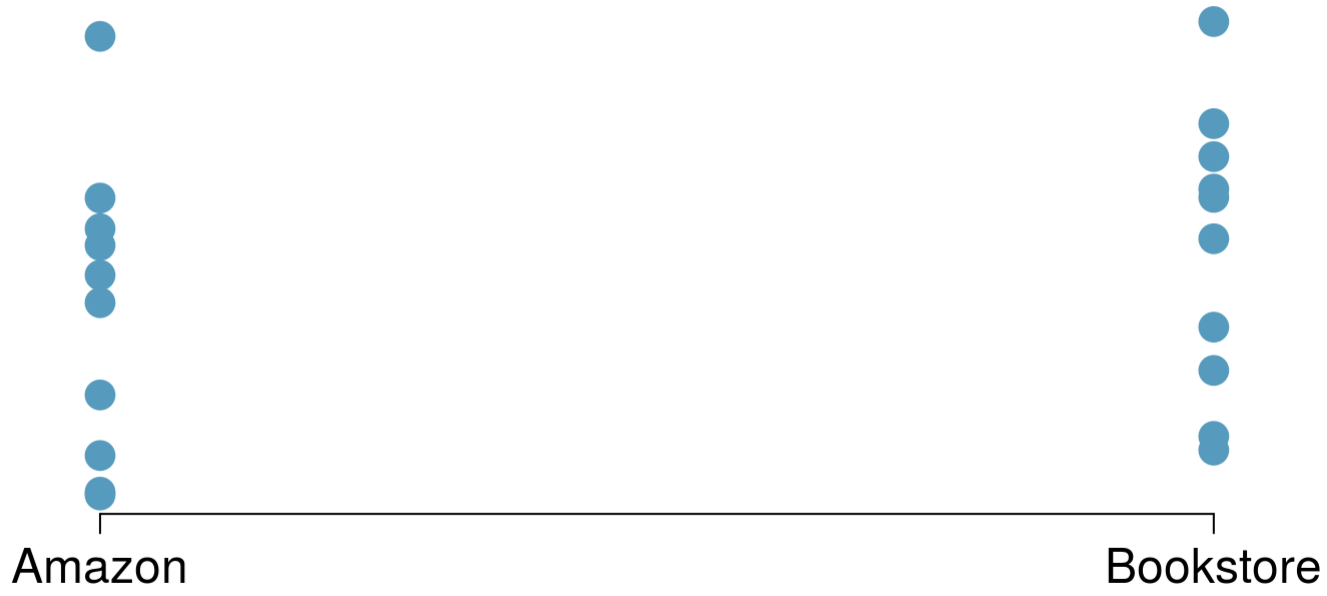
Examples of comparing means

- Is drug A better than drug B?
- Does treatment make mice smarter?
- Are textbooks cheaper online?

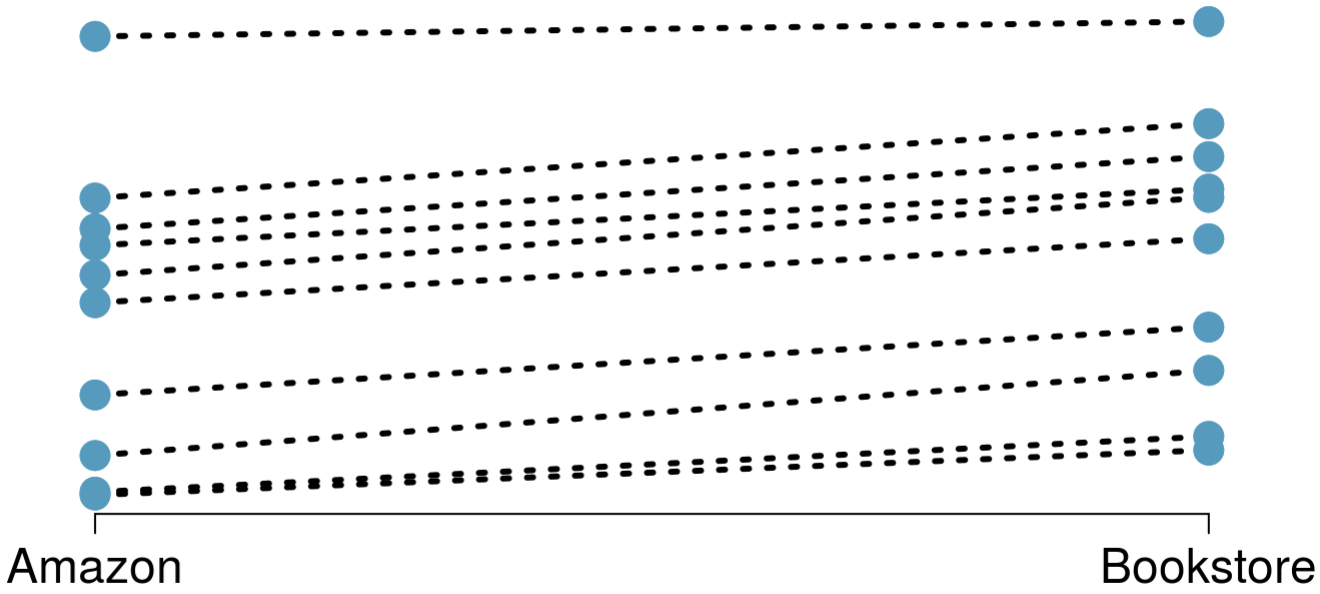
Paired vs. unpaired

- *Is drug A better than drug B?*
  - **unpaired:** give drug A to some people; give drug B to others
  - **paired:** give drug A to a person; wait a while; give drug B to same person
- *Does treatment make mice smarter?*
  - **unpaired:** give treatment to some mice; control to others
  - **paired:** measure IQ Before and After
- *Are textbooks cheaper online?*
  - **unpaired:** SRS of book prices from Amazon; SRS of book prices at the bookstore
  - **paired:** Single SRS of books. For each book, get price on Amazon and at bookstore.

Why paired makes sense



Why paired makes sense



Paired case is easy!

1. Treat **differences** as your data:

- $X_i = B_i - A_i$ difference between drug effectiveness for person i
- $X_i = \text{Before}_i - \text{After}_i$
- $X_i = B_i - A_i$ difference between bookstore and Amazon price for book title i.

1. Treat exactly like a **one-sample problem**:

$\mu = E(X_i) =$ difference in population means.

E.g., $\mu = \mu_B - \mu_A$ or $\mu_{\text{Before}} - \mu_{\text{After}}$ or $\mu_{Bookstore} - \mu_{Amazon}$.

Are one-sample assumptions met?

Knowing the price of the book at the first store tells us something about the price at the second

- Note that *within* a pair, we don't have independence
  - e.g. Before_i and After_i
- But what matters here is whether the differences are independent: $X_1, \ldots, X_n$.
  - e.g., is one person's before/after change independent of another's?
- Need either
  - $X_1, \ldots, X_n$ to be normal
  - or $n > 30$ and not too skewed.

Confidence interval for paired difference

Take $X_i = B_i - A_i$,

> A $100(1-\alpha)\%$ confidence interval for the difference $\mu_B - \mu_A$ is given by
>
> $$\bar{X}_n \pm t_{n-1,\alpha/2} S_n^X / \sqrt{n}$$

where $S_n^X = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2}$.

Test a difference with paired data

- Same as for one-sample test.
- In fact, our example of a **one-sample t-test** was actually a paired t-test in disguise!

Unpaired two-sample case

Unpaired case

- observe two independent samples:
  - $X_1, \ldots, X_{n_X}$ with mean μ_X, standard deviation σ_X
  - $Y_1, \ldots, Y_{n_Y}$ with mean μ_Y, standard deviation σ_Y
- interested in $\mu_X - \mu_Y$.
- natural statistic: $\bar{X}_{n_X} - \bar{Y}_{n_Y}$

Under usual assumptions,

$$\bar{X}_{n_X} \text{ is approximately } N\left(\mu_X, \frac{\sigma_X}{\sqrt{n_X}}\right)$$

$$\bar{Y}_{n_Y} \text{ is approximately } N\left(\mu_Y, \frac{\sigma_Y}{\sqrt{n_Y}}\right)$$

and they are independent of each other.

What can we say about $\bar{X}_{n_X} - \bar{Y}_{n_Y}$?

Fact about independent normals

If we have two independent normals,

$$A \sim N(\mu_A, \sigma_A)$$
$$B \sim N(\mu_B, \sigma_B)$$

then

$$A - B \sim N(\mu_A - \mu_B, \sqrt{\sigma_A^2 + \sigma_B^2})$$

Side note: If a and b are fixed numbers,

$$aA + bB \sim N(a\mu_A + b\mu_B, \sqrt{a^2\sigma_A^2 + b^2\sigma_B^2})$$

Back to two independent samples

$$\bar{X}_{n_X} \text{ is approximately } N\left(\mu_X, \frac{\sigma_X}{\sqrt{n_X}}\right)$$

$$\bar{Y}_{n_Y} \text{ is approximately } N\left(\mu_Y, \frac{\sigma_Y}{\sqrt{n_Y}}\right)$$

and they are independent of each other.

Therefore

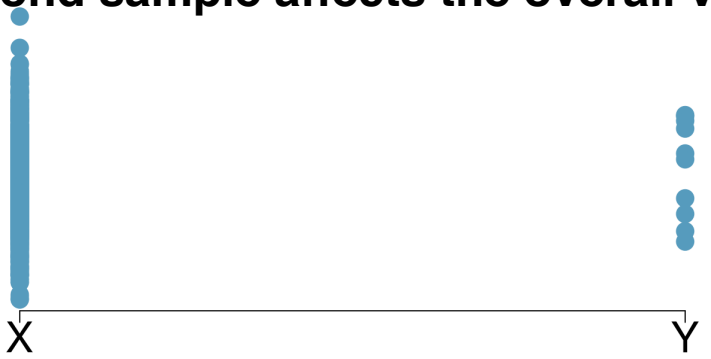
$$\bar{X}_{n_X} - \bar{Y}_{n_Y} \text{ is approx. } N\left(\mu_X - \mu_Y, \sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}\right)$$

Does variance make sense?

$$\bar{X}_{n_X} - \bar{Y}_{n_Y} \text{ is approx. } N\left(\mu_X - \mu_Y, \sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}\right)$$

What happens if $n_X = \infty$? Do we know $\mu_X - \mu_Y$ perfectly?

Even if the sample size is large for one set, variance from the second sample affects the overall variance



If variances are known

Confidence interval for $\mu_X - \mu_Y$:

$$\bar{X}_{n_X} - \bar{Y}_{n_Y} \pm z_{\alpha/2}\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}$$

based on

$$\frac{(\bar{X}_{n_X} - \bar{Y}_{n_Y}) - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}} \text{ approximately } N(0, 1).$$

Variance unknown

If $n_X > 30$ and $n_Y > 30$, then using $S^X_{n_X}$ and $S^Y_{n_Y}$ is fine:

$$\frac{(\bar{X}_{n_X} - \bar{Y}_{n_Y}) - (\mu_X - \mu_Y)}{\sqrt{\frac{[S^X_{n_X}]^2}{n_X} + \frac{[S^Y_{n_Y}]^2}{n_Y}}} \text{ approximately } N(0, 1).$$

but what if n_X or n_Y is small?

Variance unknown, small n

Unfortunately, not as simple as one-sample case

$$\frac{(\bar{X}_{n_X} - \bar{Y}_{n_Y}) - (\mu_X - \mu_Y)}{\sqrt{\frac{[S^X_{n_X}]^2}{n_X} + \frac{[S^Y_{n_Y}]^2}{n_Y}}} \text{ approximately } t_{df}$$

- this is really an approximation, not actually a t distribution even when $\bar{X}_{n_X} - \bar{Y}_{n_Y}$ is exactly normal (unless $\sigma_X = \sigma_Y$).
- complicated formula for df (R uses this).
- simpler choice (but overly conservative) is $df = \min\{n_X - 1, n_Y - 1\}$.

Example: Fuel efficiency (1973-1974)

Question: Is manual or automatic transmission more fuel efficient?

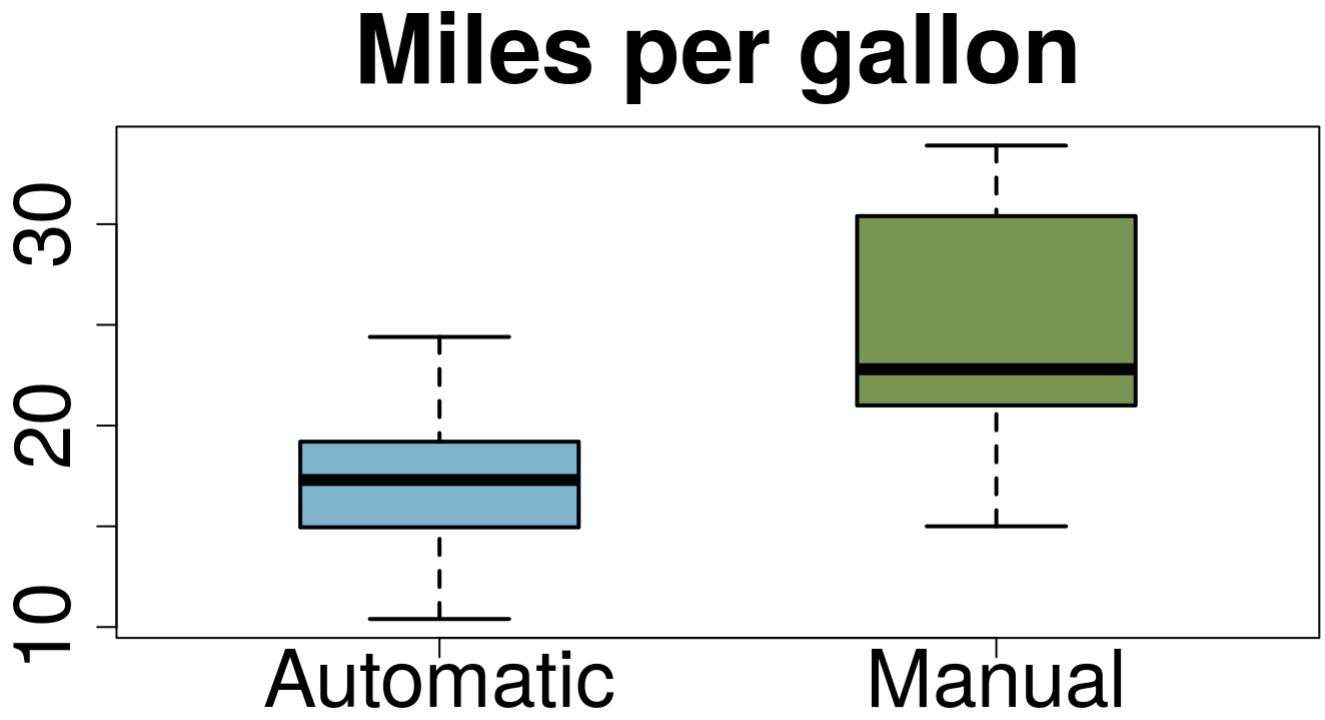
- data:
  - miles per gallon (mpg)
  - 13 manual, 19 automatic
- Hypotheses:
  - $H_0 : \mu_A = \mu_M$
  - $H_A : \mu_A \neq \mu_M$
- Paired or unpaired?

mtcars data in R (first 10 rows)

```
##                   mpg am
## Mazda RX4         21.0  1
## Mazda RX4 Wag     21.0  1
## Datsun 710        22.8  1
## Hornet 4 Drive    21.4  0
## Hornet Sportabout 18.7  0
## Valiant           18.1  0
## Duster 360        14.3  0
## Merc 240D         24.4  0
## Merc 230          22.8  0
## Merc 280          19.2  0
```

Always plot first

```
boxplot(mpg ~ am, data = mtcars, names = c("Automatic", "Manual"),
        main="Miles per gallon")
```



Check assumptions

- Need $\bar{X}_{n_X}$ approximately normal…
  - $n_X = 19$… would need data distribution normal
  - independent?
- Same for $\bar{Y}_{n_Y}$.
- And need the two samples independent (since not paired)

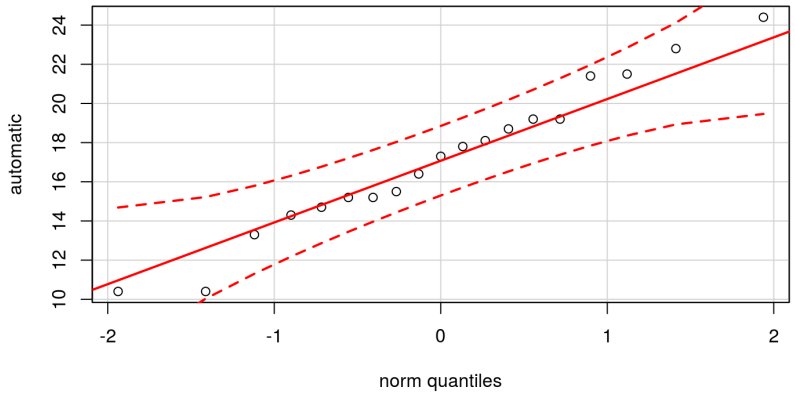
In R

```
automatic <- subset(mtcars, am == 0)$mpg
manual <- subset(mtcars, am == 1)$mpg
library(car) # for qqPlot function
```

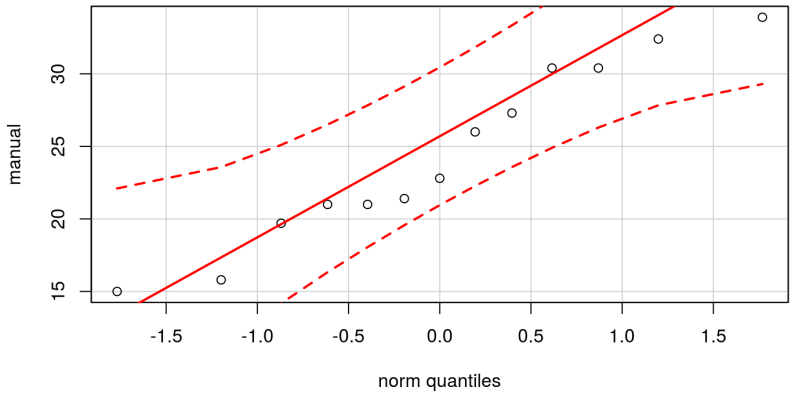
- package `car` has nothing to do with cars (stands for “Companion to Applied Regression”)

Approximately normal?

```
qqPlot(automatic)
```




```
qqPlot(manual)
```



Doing test

```
t.test(x = manual, y = automatic)
```

```
##
##  Welch Two Sample t-test
##
## data:  manual and automatic
## t = 3.7671, df = 18.332, p-value = 0.001374
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   3.209684 11.280194
## sample estimates:
## mean of x mean of y
##  24.39231  17.14737
```

Example: Fertilized Tomatoes

Question: Does a new fertilizer improve tomato yield?

- 30 plots of tomato plants, 10 plants per plot.
- Treat random 5 plants with new fertilizer in each plot
- Measure yield (in kg) from each
- Is this paired or unpaired?

In R

Can either use

```
t.test(x, y, paired = TRUE)
```

or

```
t.test(x - y)
```

Pairing (more on this in 6020)

- aka, **matching** reduces uncontrolled sources of variability (e.g., sunlight).
- can be useful in observational studies to control for *confounding* (variables associated with both variables of interest)
- example of **blocking** (a key principle of experimental design) - create blocks of relatively homogenous units and then assess treatment effect by using "within-block" differences.
- pairing is special case of more general problem: **repeated measurements** (two or more measurements of same block/unit)
  - multiple treatments per block
  - or longitudinal data on each subject

A look back at paired vs. unpaired

- the test statistics both are of the form

$$\frac{\bar{X}_{n_X} - \bar{Y}_{n_Y}}{\widehat{SE}(\bar{X}_{n_X} - \bar{Y}_{n_Y})}$$

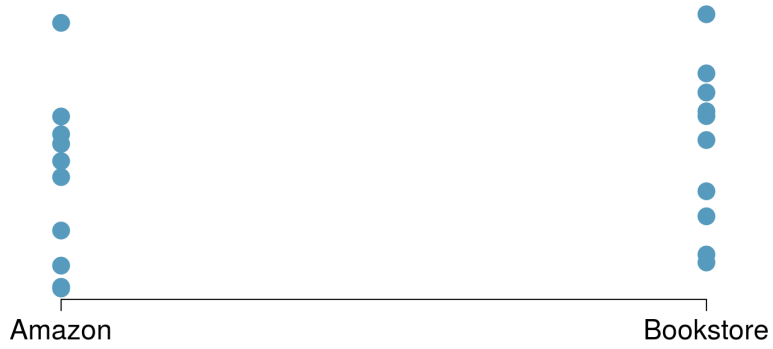
  where $\widehat{SE}(\bar{X}_{n_X} - \bar{Y}_{n_Y})$ is an estimate of the standard deviation of numerator

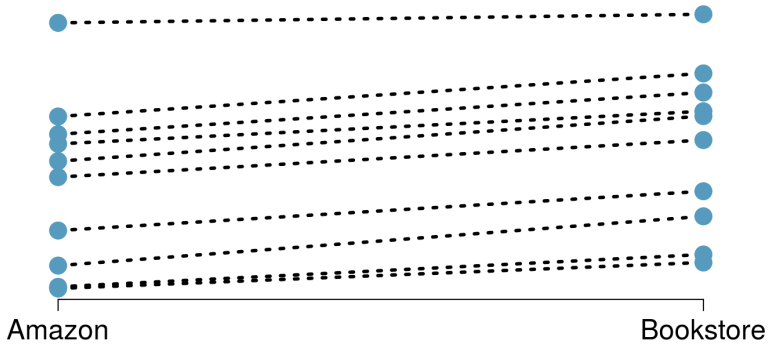
- **really?** recall in paired case we use $D_i = X_i - Y_i$ and look at

$$\bar{D}_n = \frac{1}{n}\sum_{i=1}^{n} D_i = \frac{1}{n}\sum_{i=1}^{n}(X_i - Y_i)$$

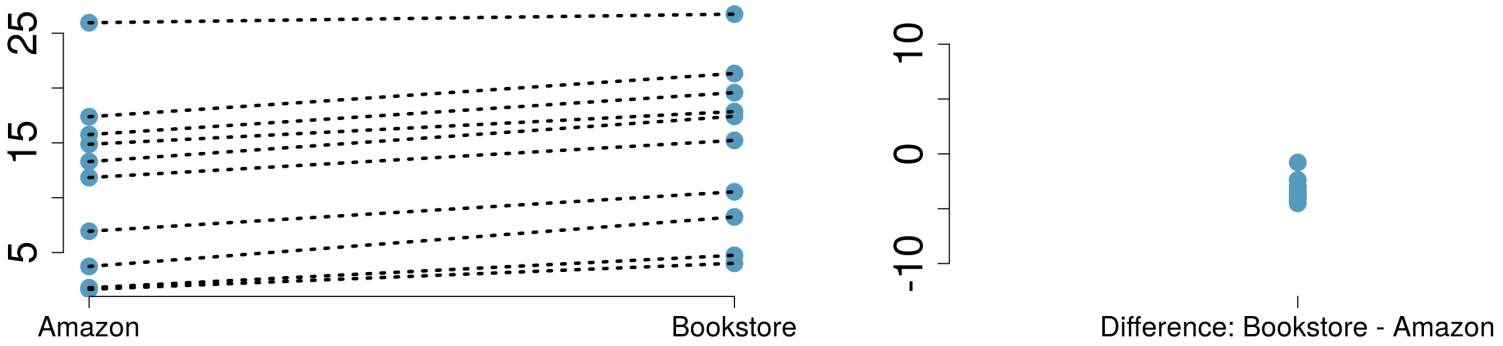
$$= \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) - \left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \bar{X}_n - \bar{Y}_n$$

Back to that picture





Variance of the differences



So then what's different?

Two differences:

1. **Denominator:** $\widehat{SE}(\bar{X}_{n_X} - \bar{Y}_{n_Y})$
2. **Distribution** of test statistic

1) Denominator

What is the standard deviation of $\bar{X}_{n_X} - \bar{Y}_{n_Y}$?

- *paired:* depends on standard deviation **of difference**:

$$\frac{\sigma_D}{\sqrt{n}}$$

- *unpaired:* depends on how much variation within each group

$$\sqrt{\sigma_X^2/n_X + \sigma_Y^2/n_Y}$$

Assuming X_i and Y_i are positively correlated, $D_i = X_i - Y_i$ may vary much less than if they are uncorrelated.

2) Distribution

- *paired:* t_{n-1} distribution under H_0 (for n pairs)
- *unpaired:* t_{df} for some df depending on n_X and n_Y.