T-tests: Assumptions

Research Methods for Human Inquiry
Andrew Perfors

Assumptions

- All t-tests:
  - The population distribution* of each group is normal
  - The data are independent except in those respects that the test specifies (e.g. paired samples t-test implies a very specific kind of relationship between data from the same person)

- Student t-test:
  - The two groups are drawn from populations with the same variance (homogeneity of variance assumption)

Technically the assumption is about the normality of the residuals but for t-tests this amounts to the same thing

Checking the normality assumption

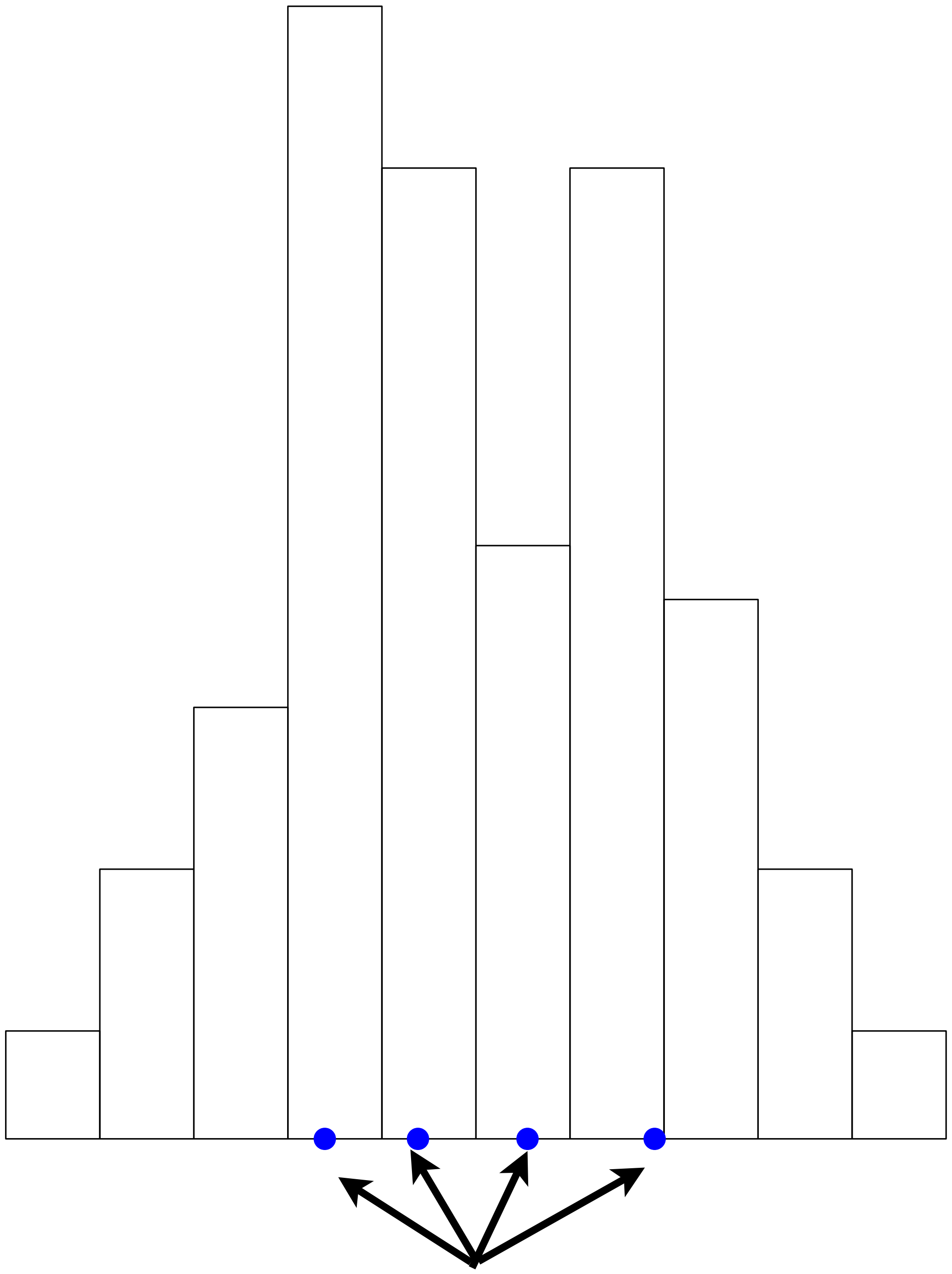
Important Note: There are some good arguments I've heard about how it isn't necessary to check the normality assumption: not only are most tests fairly robust to violations of normality, but most tests of normality aren't very trustworthy because they will always indicate that something is not normal with large enough sample size. I myself usually check but only worry if things are *very* obviously not normal. I'm teaching you this because many people care so I want you to know how to do it, and because it is good to know what to do if there are major violations of normality. For assessments in this subject, I'll be expecting that you follow the guidelines here.

The LMS has a few links going into more detail if you're curious (optional)

Quantile-quantile (QQ) plots

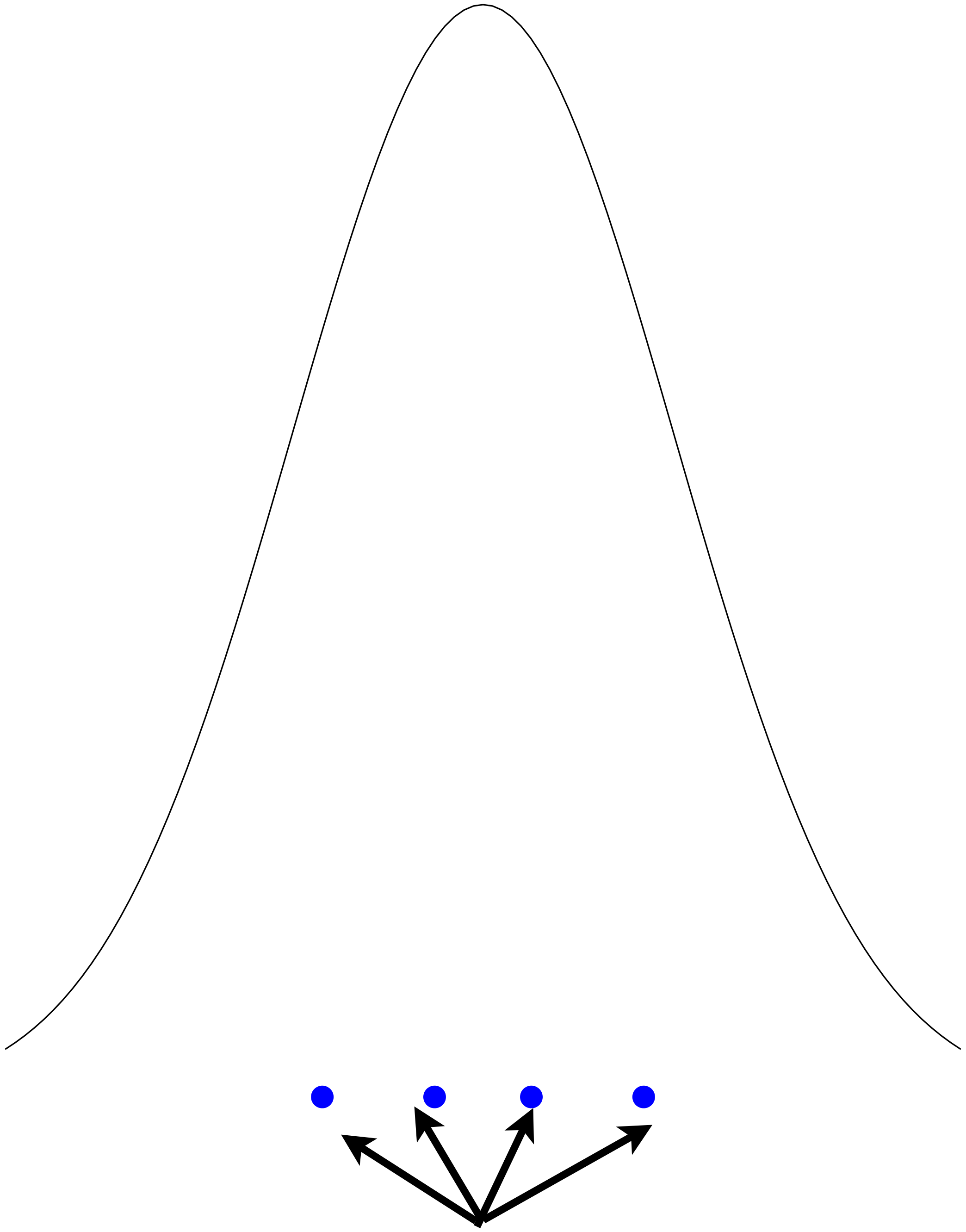
- Scatterplot of the actual quantiles of the observed data against the theoretical quantiles of the normal distribution
- If the data are truly normally distributed, you'd expect the quantiles to be identical, giving you a nice straight line…

The observed data



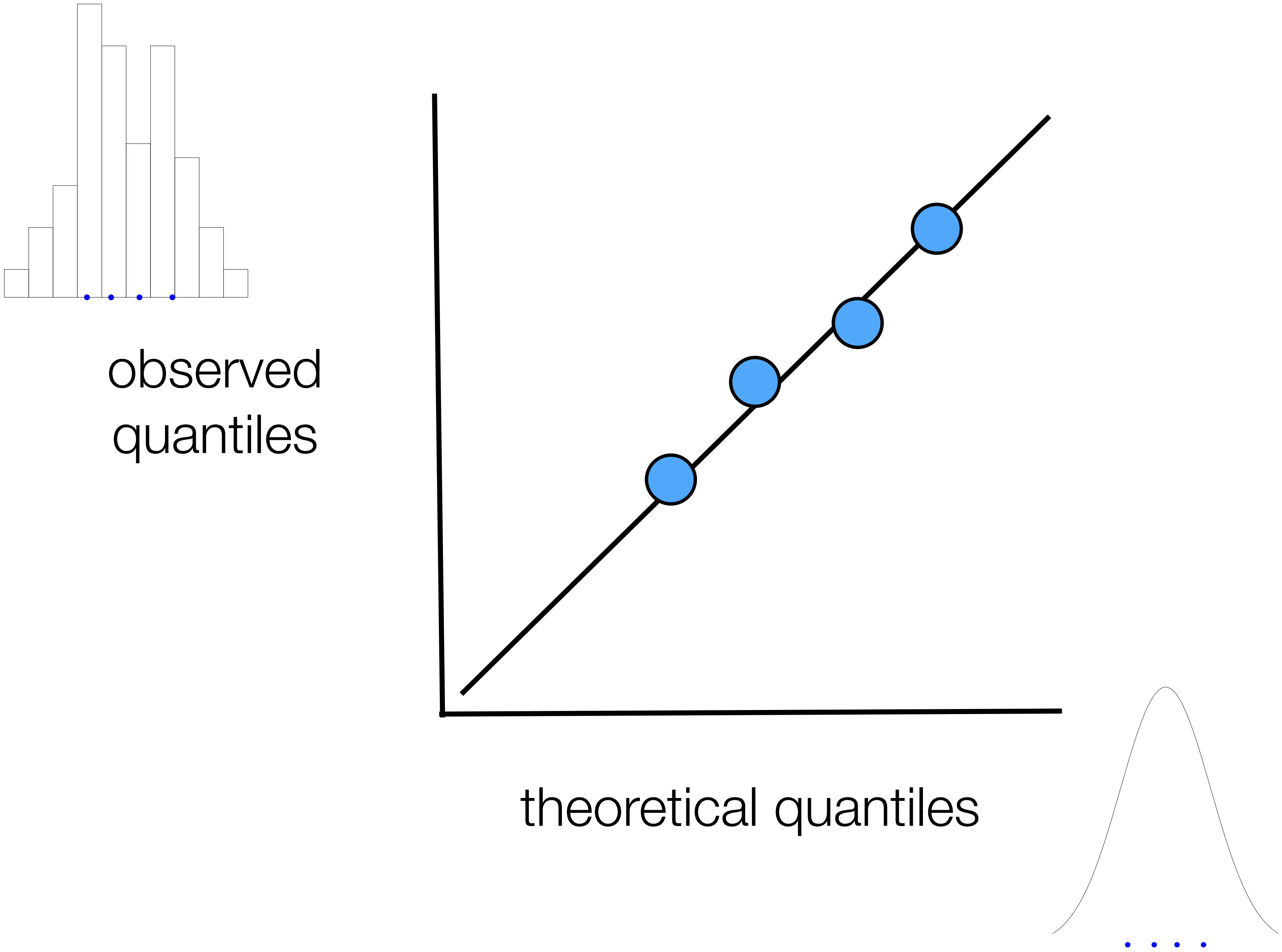
20th, 40th, 60th and 80th percentiles

Normal distribution with the same mean and standard deviation as the data



theoretical values for the 20th, 40th, 60th and 80th percentiles area almost identical to the empirically observed ones…

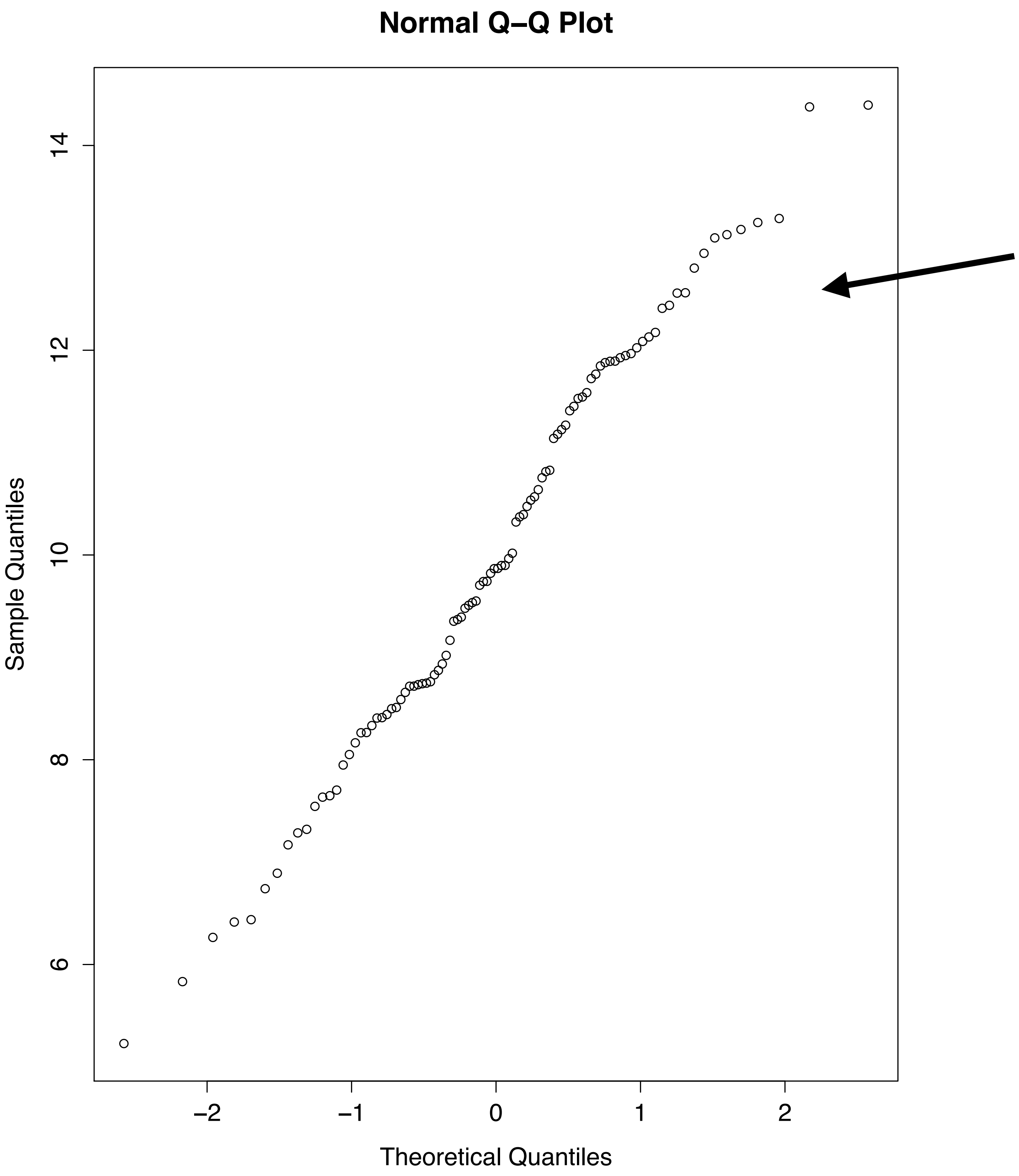
A QQ plot showing only 4 quantiles...



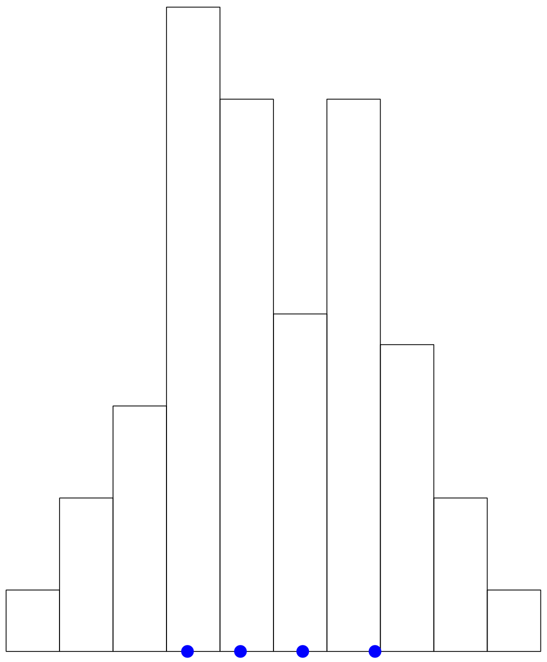
observed quantiles

theoretical quantiles

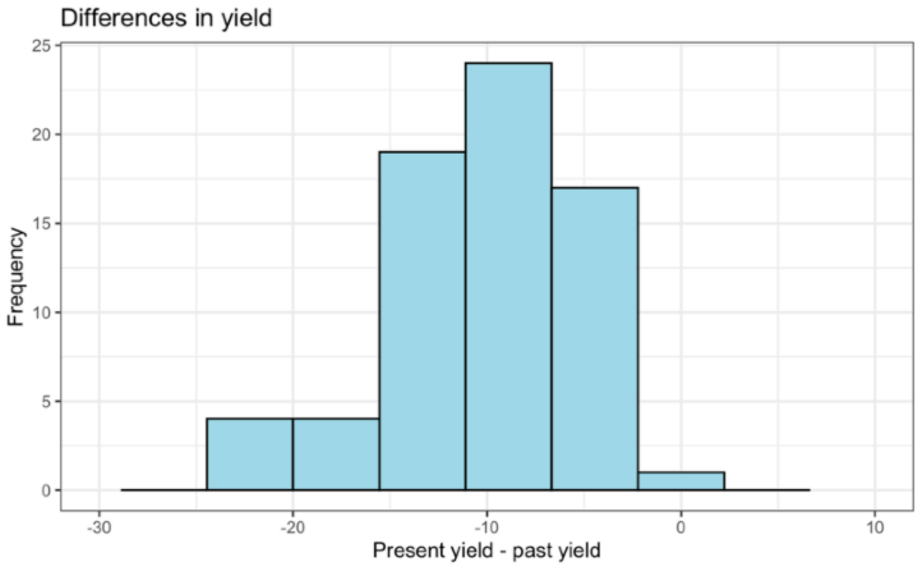
The real QQ plot, with every data point in the sample included...

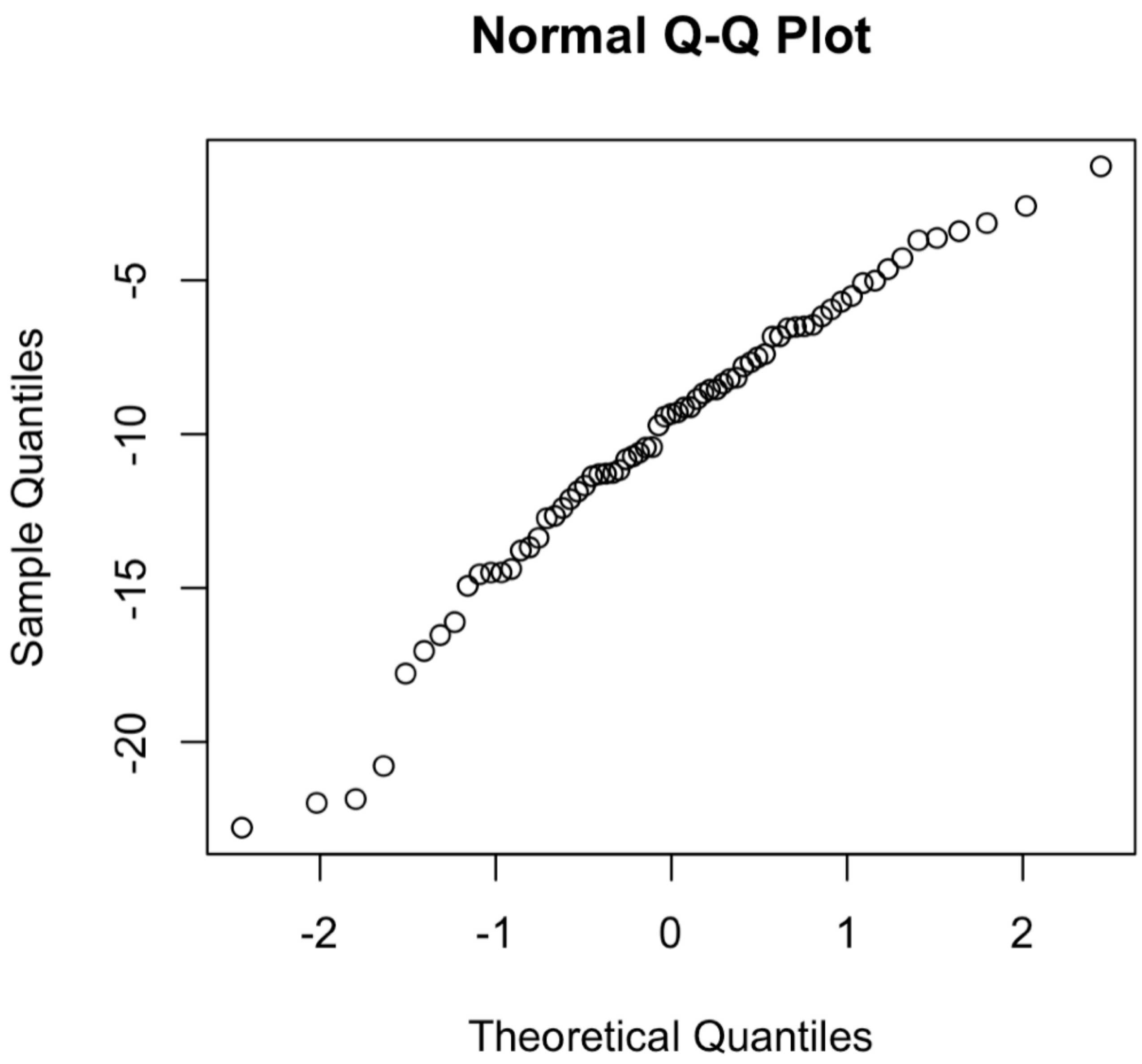


Straight line implies normality



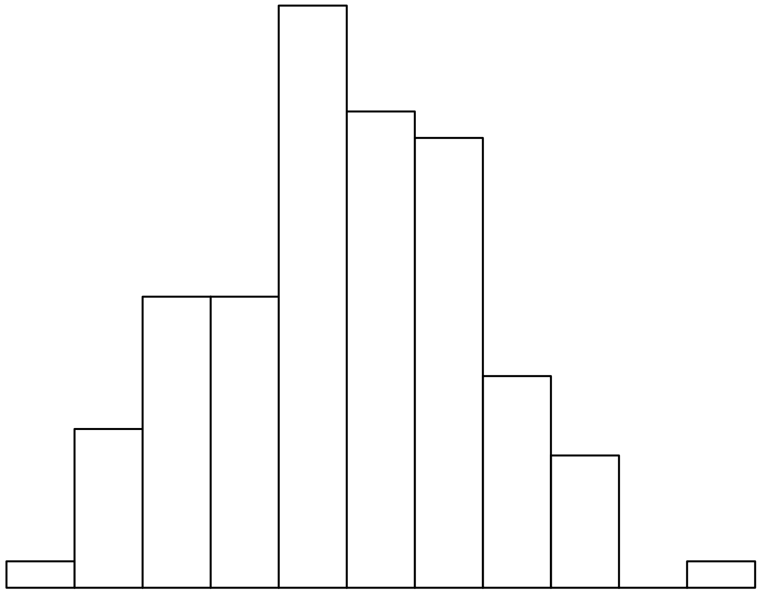
```
qqnorm(dch$diff)
```

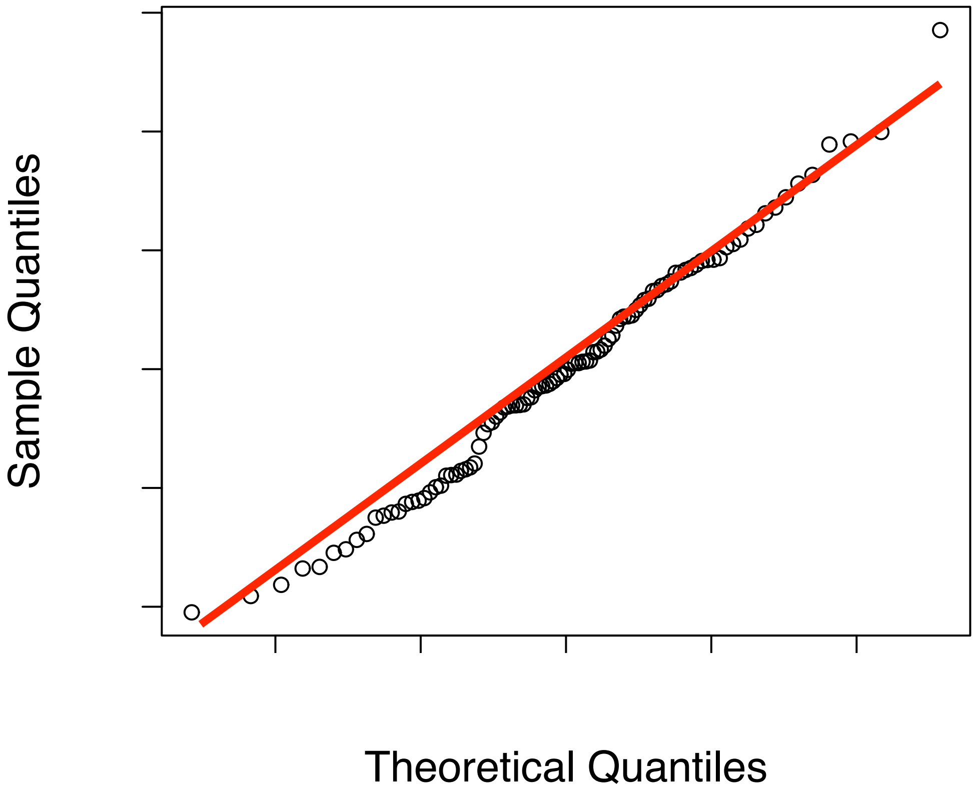




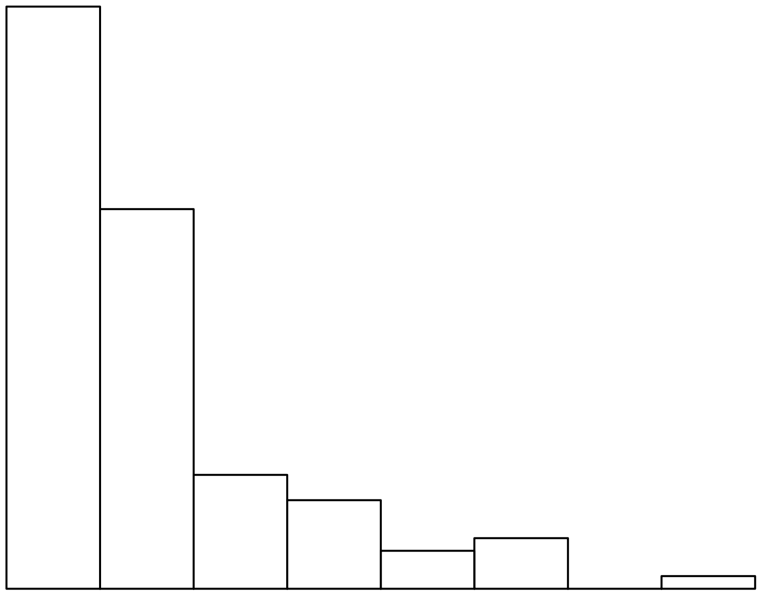
Use the `qqnorm()` function to draw the plot

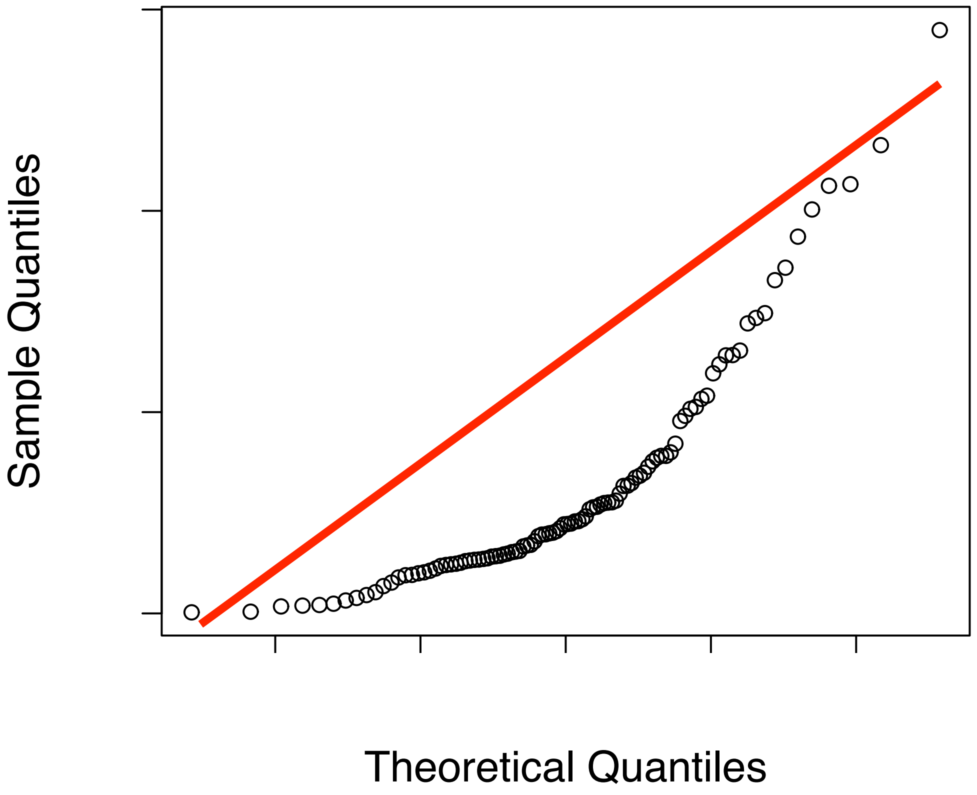
If your data are normal…





If your data are skewed…





The Shapiro-Wilk test

- A way of quantifying departures from normality
  - I won't discuss the details of how it works.
  - The test statistic is called W. Values of W less than 1 imply deviations from normality
  - The R command is `shapiro.test()`

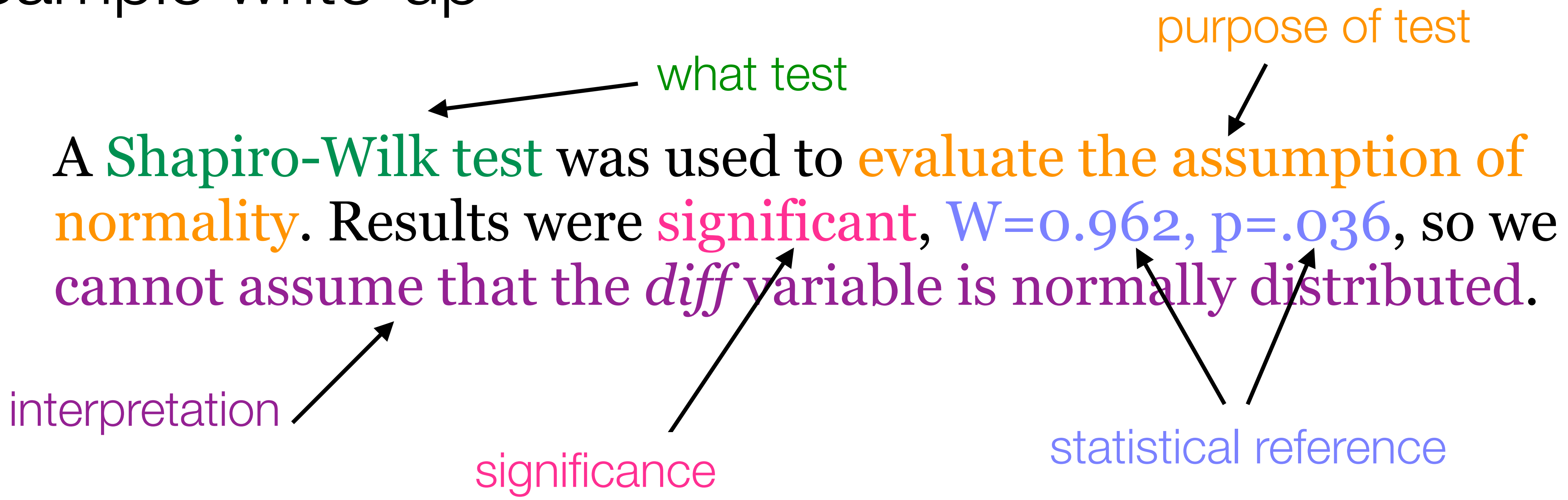
```
> shapiro.test(dch$diff)

	Shapiro-Wilk normality test

data:  dch$diff
W = 0.96233, p-value = 0.03576
```

null hypothesis is that the data are normally distributed

p-value less than 0.05 means you reject the null: data are not normal!

The Shapiro-Wilk test

Sample write-up



A Shapiro-Wilk test was used to evaluate the assumption of normality. Results were significant, W=0.962, p=.036, so we cannot assume that the *diff* variable is normally distributed.

```
> shapiro.test(dch$diff)

    Shapiro-Wilk normality test

data:  dch$diff
W = 0.96233, p-value = 0.03576
```

null hypothesis is that the data are normally distributed

p-value less than 0.05 means you reject the null: data are not normal!

The Shapiro-Wilk test

- **Important**: Frequently significant if the sample size is too large (above 50!) even if the distribution is normal enough

- If this happens to you and your sample size is over 50, take a look at the Q-Q plot and histogram. If they both look fine, you're probably safe to run a traditional t-test

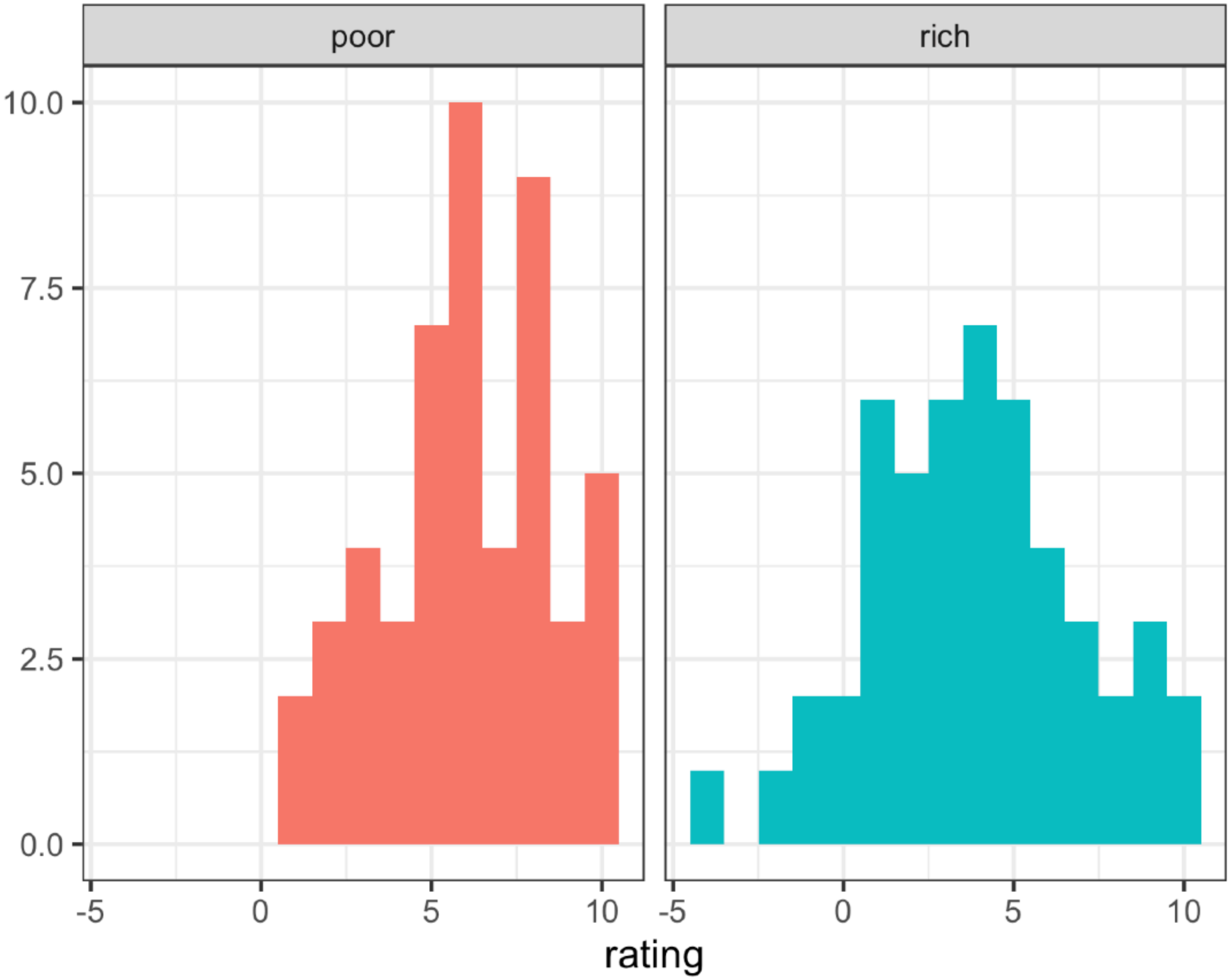
- If you're at all unsure, it won't hurt to NOT to a traditional test. You'll lose a bit of power but it's safest.

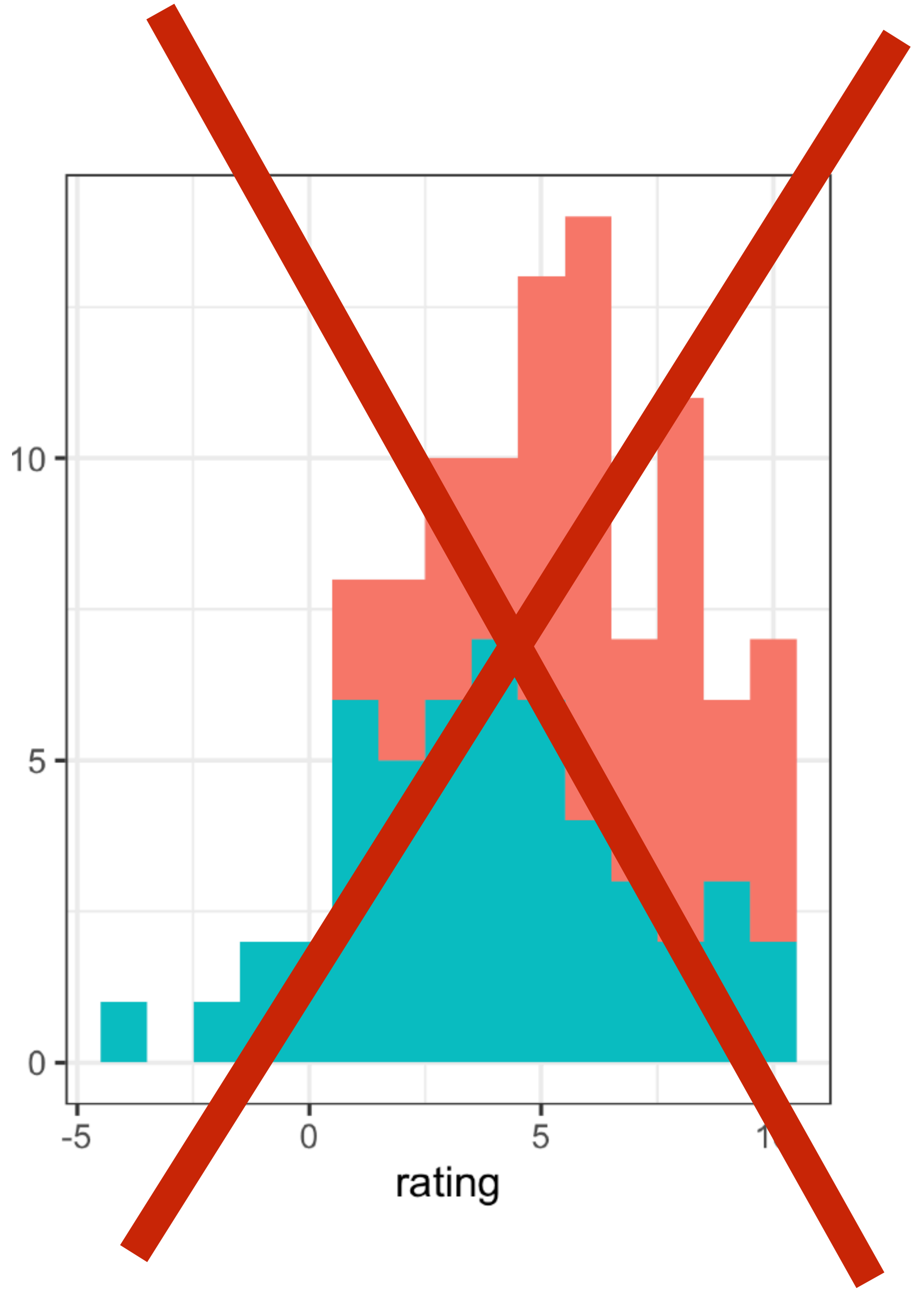
- There's are other tests of normality (e.g., Kolmogorov-Smirnov) but all have this issue to some extent.

- For RMHI, use the Shapiro-Wilk and treat p<.05 as not normal. I'm mentioning the nuance for the real world!

Important note for independent sample t-test

- If you have **multiple groups**, you need to check the normality of each one separately





Important note for independent sample t-test

- If you have **multiple groups**, you need to check the normality of each one separately

```
> shapiro.test(de$rating[de$ses=="rich"])

	Shapiro-Wilk normality test

data:  de$rating[de$ses == "rich"]
W = 0.98295, p-value = 0.6813

> shapiro.test(de$rating[de$ses=="poor"])

	Shapiro-Wilk normality test

data:  de$rating[de$ses == "poor"]
W = 0.95654, p-value = 0.0637
```

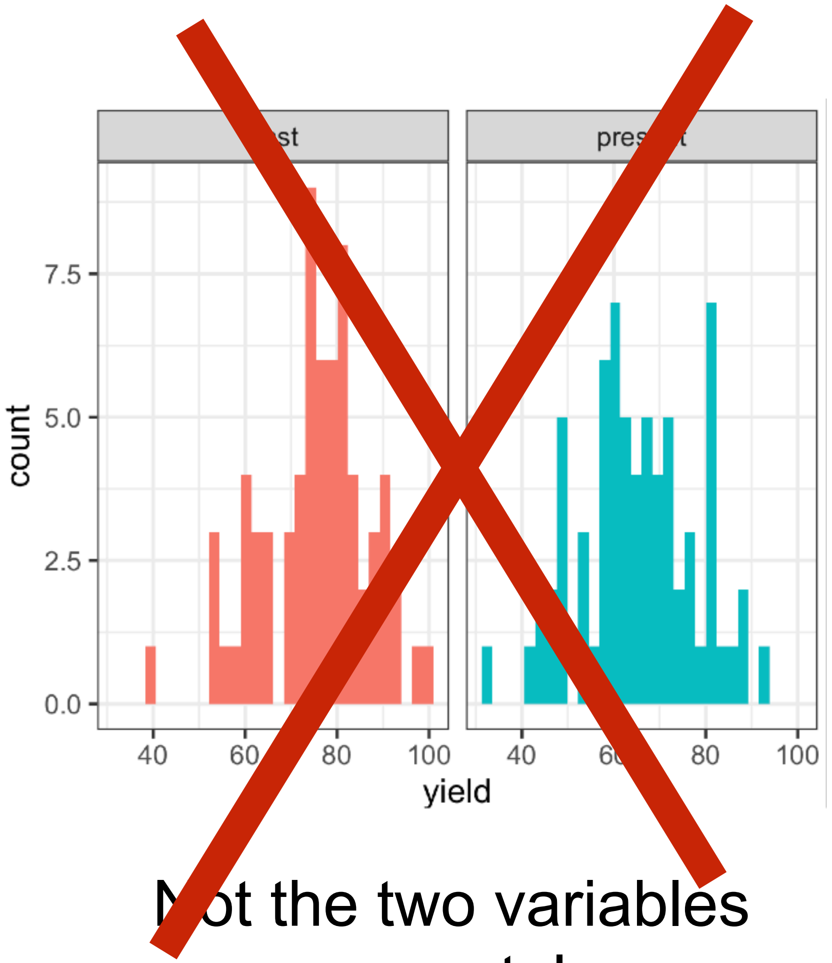
null hypothesis is that the data are normally distributed

p-value greater than 0.05 means you fail to reject the null: ok to assume the data are normal!

note: all groups must be normal for this assumption to be met

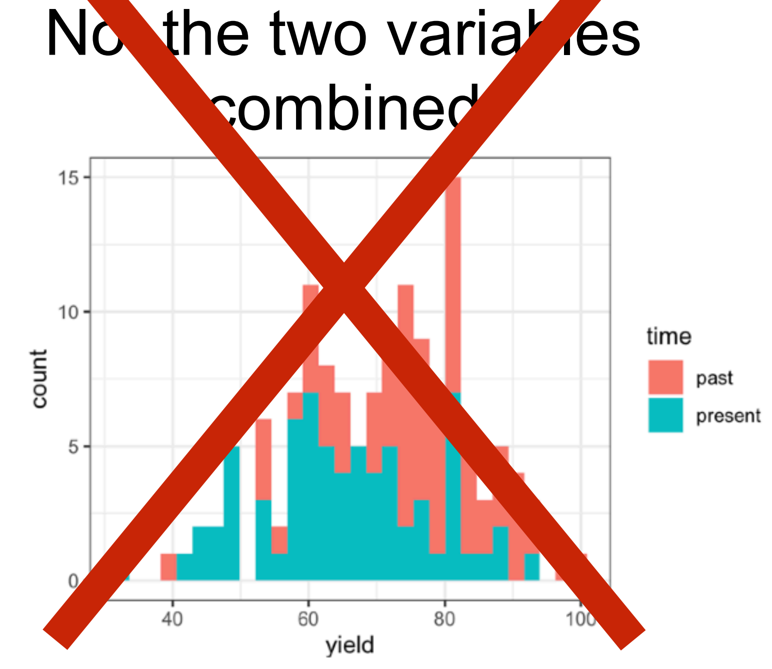
Important note for paired-sample t-test

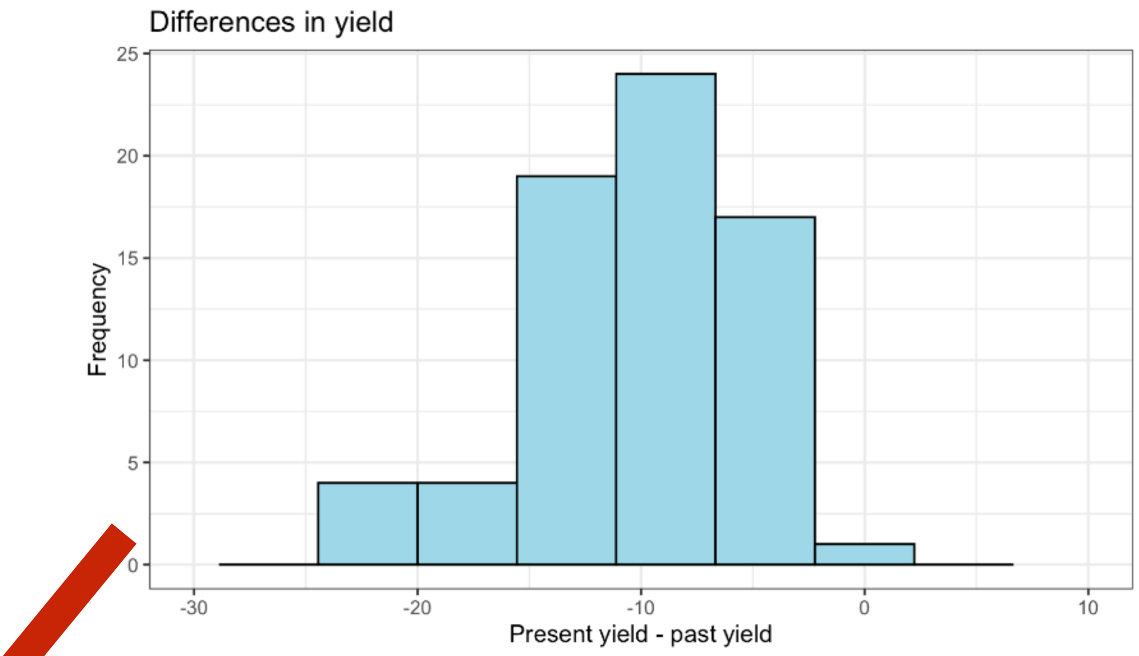
- If you are doing a paired-sample t-test, you need to check the **normality of the *difference* variable**
- This is because a paired-sample t-test is the same as a one-sample t-test, so the difference variable is the relevant one



Not the two variables separately

Not the two variables combined





Check the normality of the difference variable

What to do if the data are not normal?

Non-parametric tests

- Non-parametric tests avoid making assumptions about the shape of the distribution (e.g., normal)
  - Usually not as powerful (i.e., higher Type II error) than the corresponding parametric tests
  - But sometimes you just have to use them.

- Non-parametric version of the t-test?
  - Called the Wilcoxon test.
  - *Learning Statistics with R* gives you more details if you need the info. I'll give you a basic idea of the independent-samples version and show you how to run both.

Independent samples Wilcoxon test

The test statistic W counts the number of times a score from group A is larger than a score from group B

Take the scores from both groups…

Organise them into a table

group B

group A

	14.5	10.4	12.4	11.7	13
6.4					
10.7		●			
11.9		●		●	
7.3					
10					

W=3

What to expect based on the null hypothesis is complex, but basically about half of the possibilities should be larger if H0 is true

Independent samples Wilcoxon: In R

```
> wilcox.test( rating ~ ses, data = de )

	Wilcoxon rank sum test with continuity correction

data:  rating by ses
W = 1778.5, p-value = 0.0002537
alternative hypothesis: true location shift is not equal to 0
```

Test statistic and p-value

Alternative hypothesis says there is a “shift” or “difference” between groups

Independent samples Wilcoxon: In R

```
> wilcox.test( rating ~ ses, data = de )

	Wilcoxon rank sum test with continuity correction

data:  rating by ses
W = 1778.5, p-value = 0.0002537
alternative hypothesis: true location shift is not equal to 0
```

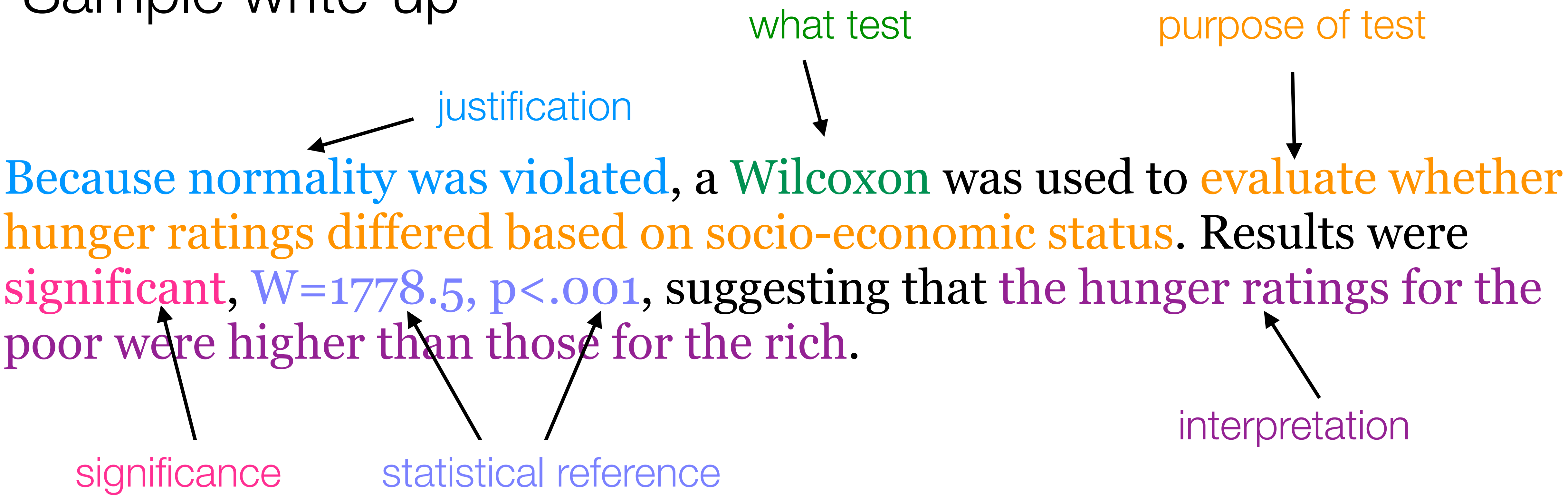
```
cannot compute exact p-value with ties
```

Sometimes you'll see this: the Wilcoxon test doesn't like ties and R resorts to an approximation to compute the p-value. The issue is fixable, but we won't worry about it here. It's not a big deal unless there are a lot of ties (i.e., items where the scores are the same for both groups).

Independent samples Wilcoxon: In R

```
> wilcox.test( rating ~ ses, data = de )

	Wilcoxon rank sum test with continuity correction

data:  rating by ses
W = 1778.5, p-value = 0.0002537
alternative hypothesis: true location shift is not equal to 0
```

Sample write-up



Because normality was violated, a Wilcoxon was used to evaluate whether hunger ratings differed based on socio-economic status. Results were significant, W=1778.5, p<.001, suggesting that the hunger ratings for the poor were higher than those for the rich.

Paired samples Wilcoxon: In R

```
> wilcox.test( yield ~ time, data = dc_long2, paired=TRUE )

	Wilcoxon signed rank test with continuity correction

data:  yield by time
V = 2415, p-value = 5.331e-13
alternative hypothesis: true location shift is not equal to 0
```

all the other output is the same!

Effect size for Wilcoxon

Cohens D is defined assuming normality holds. If it does not, need to compute the Wilcoxon effect size (r) instead.

Test statistic converted to Z score

$$r = \frac{Z}{\sqrt{N}}$$

square root of sample size

Similar interpretation as Cohens D if you want that

r	rough interpretation
0.1-0.3	small
0.3-0.5	medium
> 0.5	large

Effect size for Wilcoxon

```
> install.packages("rstatix")
> library(rstatix)
```

For two independent samples

```
> wilcox_effsize(rating~ses,data=de)
  .y.    group1 group2 effsize    n1    n2 magnitude
* <chr>  <chr>  <chr>    <dbl> <int> <int> <ord>
1 rating poor   rich     0.366    50    50 moderate
```

For paired samples

```
> wilcox_effsize(yield~time,data=dc_long2,paired=TRUE)
  .y.   group1 group2  effsize    n1    n2 magnitude
* <chr> <chr>  <chr>     <dbl> <int> <int> <ord>
1 yield past   present   0.869    69    69 large
```

Exercises are in w7day2exercises.Rmd