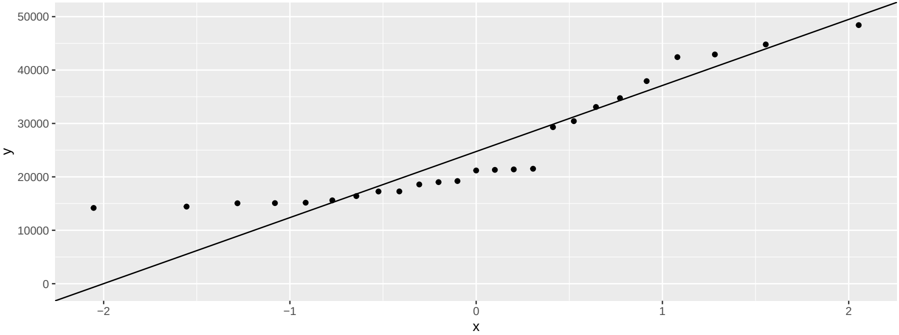
Assumptions

Assumptions

- The t procedures we have seen so far come with assumption of normally-distributed data
- but how much does that normality matter?
- Central Limit Theorem says that sampling distribution of sample mean is “approximately normal” if sample size is “large”.
- Hence same applies to difference of two sample means.
- How to use this in practice? Draw a picture and make a call about whether sample size large enough.

Blue Jays attendances

```
ggplot(jays, aes(sample = attendance)) +
  stat_qq() + stat_qq_line()
```



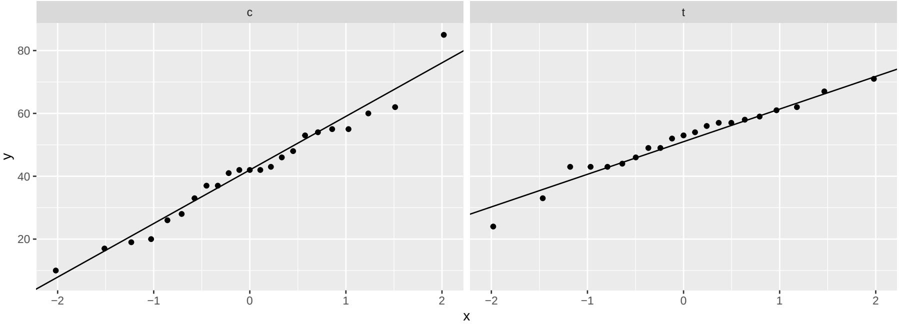
Comments

- Distribution of attendances somewhat skewed to the right (because of the short lower tail and the sort-of curve)
- Sample size $n = 25$ is reasonably large in Central Limit Theorem terms
- Use of t *may* be OK here despite skewed shape.

Learning to read

- Make normal quantile plots, one for each sample:

```
ggplot(kids, aes(sample = score)) +
  stat_qq() + stat_qq_line() +
  facet_wrap(~ group)
```



Comments

- with sample sizes over 20 in each group, these are easily normal enough to use a t-test.
- the (sampling distribution of the) difference between two sample means tends to have a more normal distribution than either sample mean individually, so that two-sample t tends to be better than you'd guess.

Pain relief

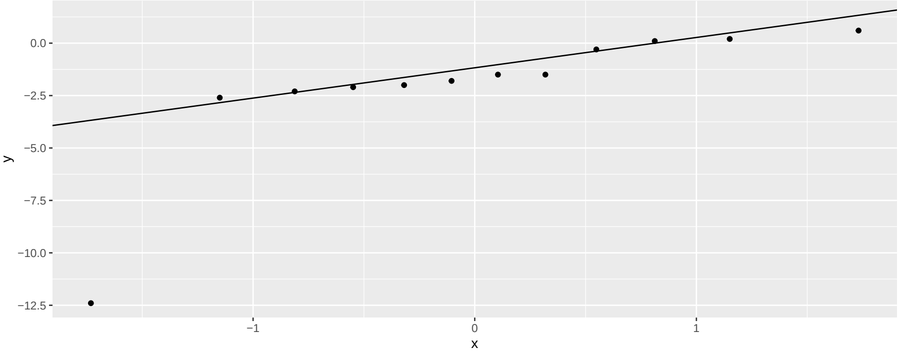
- With matched pairs, assumption is of normality of *differences*, so work those out first:

```
pain %>% mutate(diff = druga - drugb) -> pain
pain
```

```
# A tibble: 12 x 4
   subject druga drugb    diff
     <dbl> <dbl> <dbl>   <dbl>
 1       1   2     3.5  -1.5
 2       2   3.6   5.7  -2.1
 3       3   2.6   2.9  -0.300
 4       4   2.6   2.4   0.200
 5       5   7.3   9.9  -2.6
 6       6   3.4   3.3   0.100
 7       7  14.9  16.7  -1.80
 8       8   6.6   6     0.600
 9       9   2.3   3.8  -1.5
10      10   2     4    -2
11      11   6.8   9.1  -2.3
12      12   8.5  20.9 -12.4
```

Normality of differences

```
ggplot(pain,aes(sample=diff)) + stat_qq() + stat_qq_line()
```



Comments

- This is very non-normal (the low outlier)
- The sample size of $n = 12$ is not large
- We should have concerns about our matched pairs t-test.

Doing things properly

- The right way to use a t procedure:
  - draw a graph of our data (one of the standard graphs, or normal quantile plot)
  - use the graph to assess sufficient normality given the sample size
  - for a two-sample test, assess equality of spreads (boxplot easier for this)
  - if necessary, express our doubts about the t procedure (for now), or do a better test (later).

Looking ahead

- Looking at a normal quantile plot and assessing it with the sample size seems rather arbitrary. Can we do better? (Yes: using the bootstrap, later.)
- What to do if the t procedure is not to be trusted? Use a different test (later):
  - one sample: sign test
  - two samples: Mood's median test
  - matched pairs: sign test on differences.
- If you have heard about the signed rank or rank sum tests: they come with extra assumptions that are usually not satisfied if normality fails.