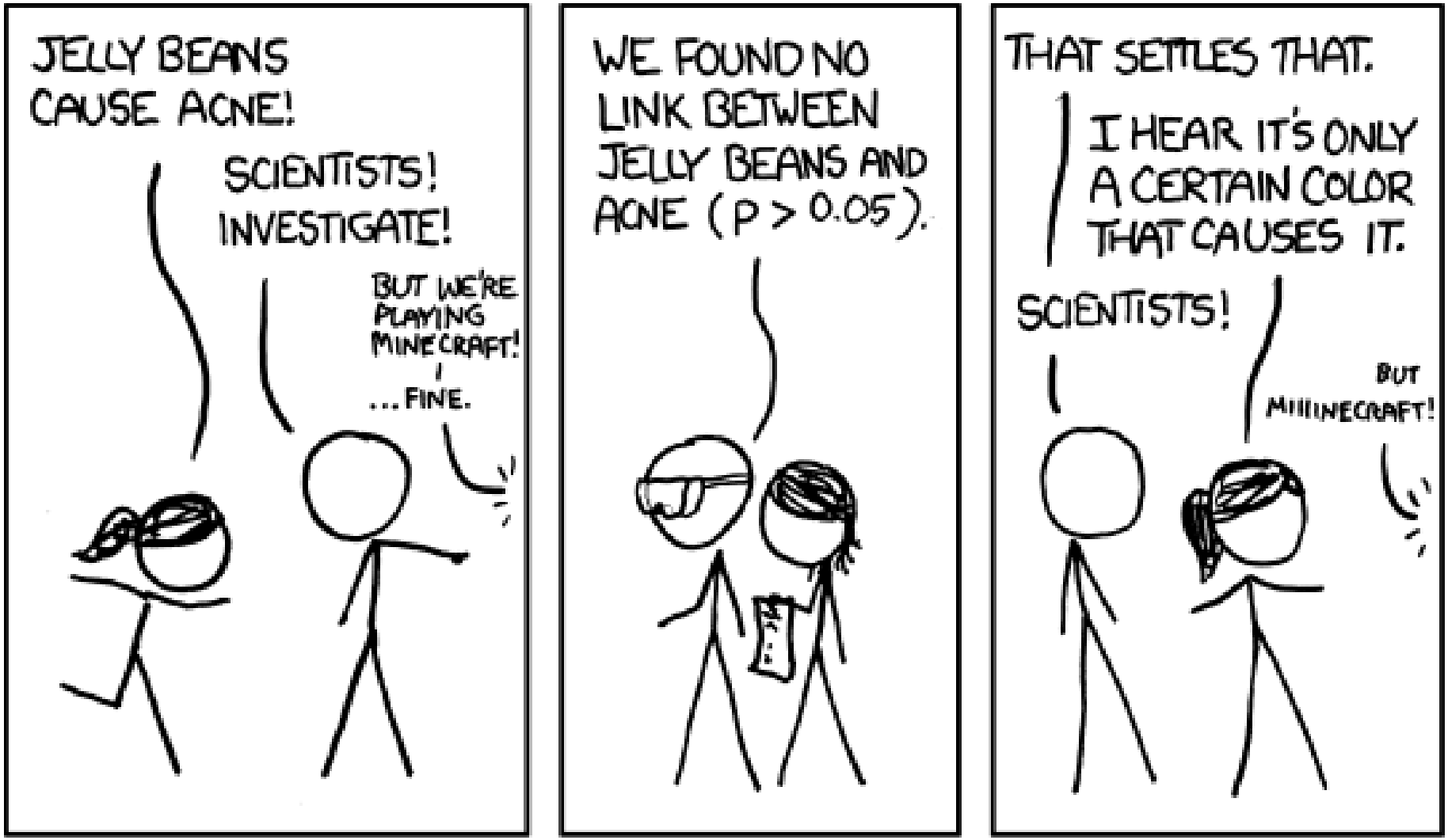
431 Class 18

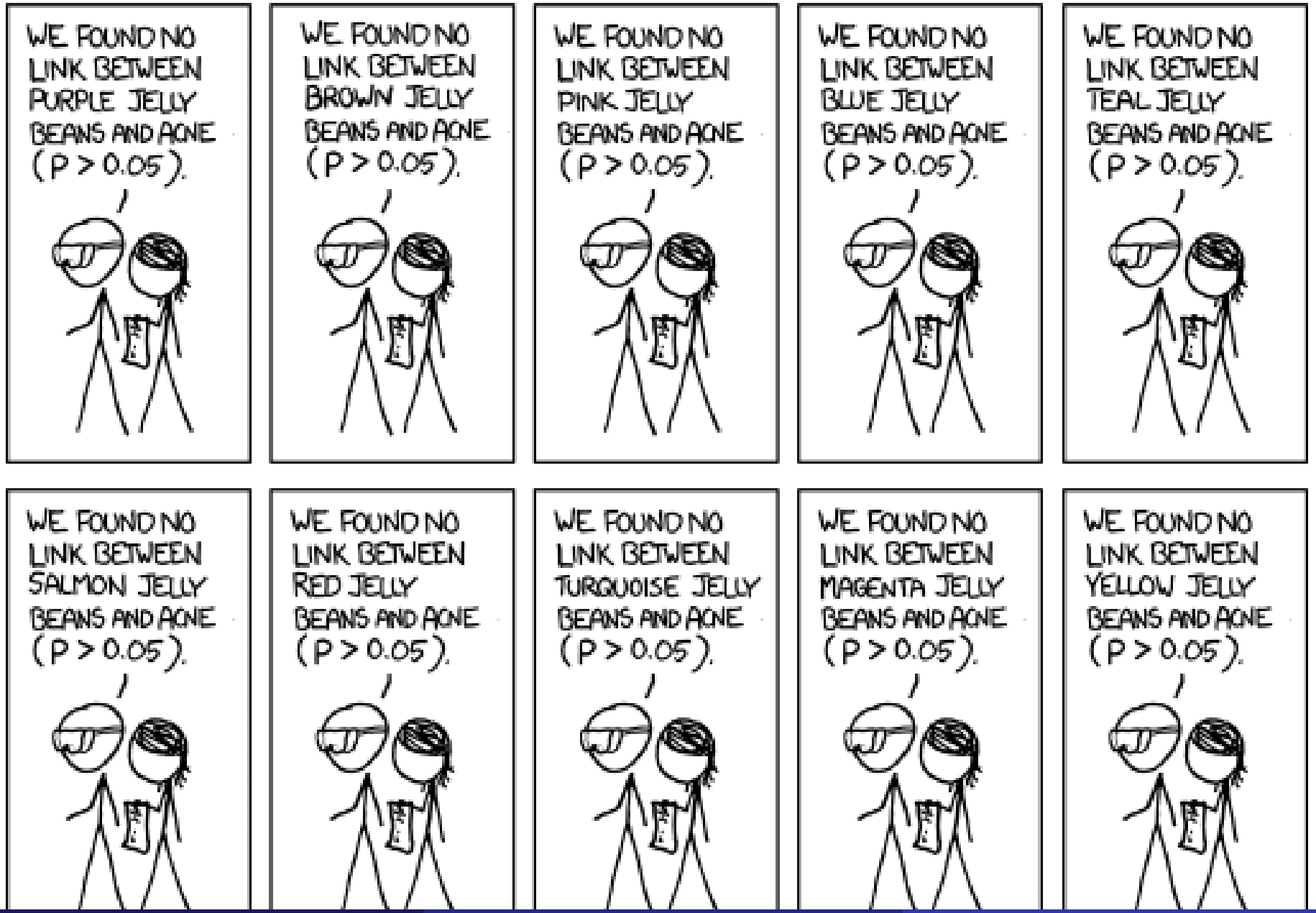
thomaselove.github.io/431

2020-10-27

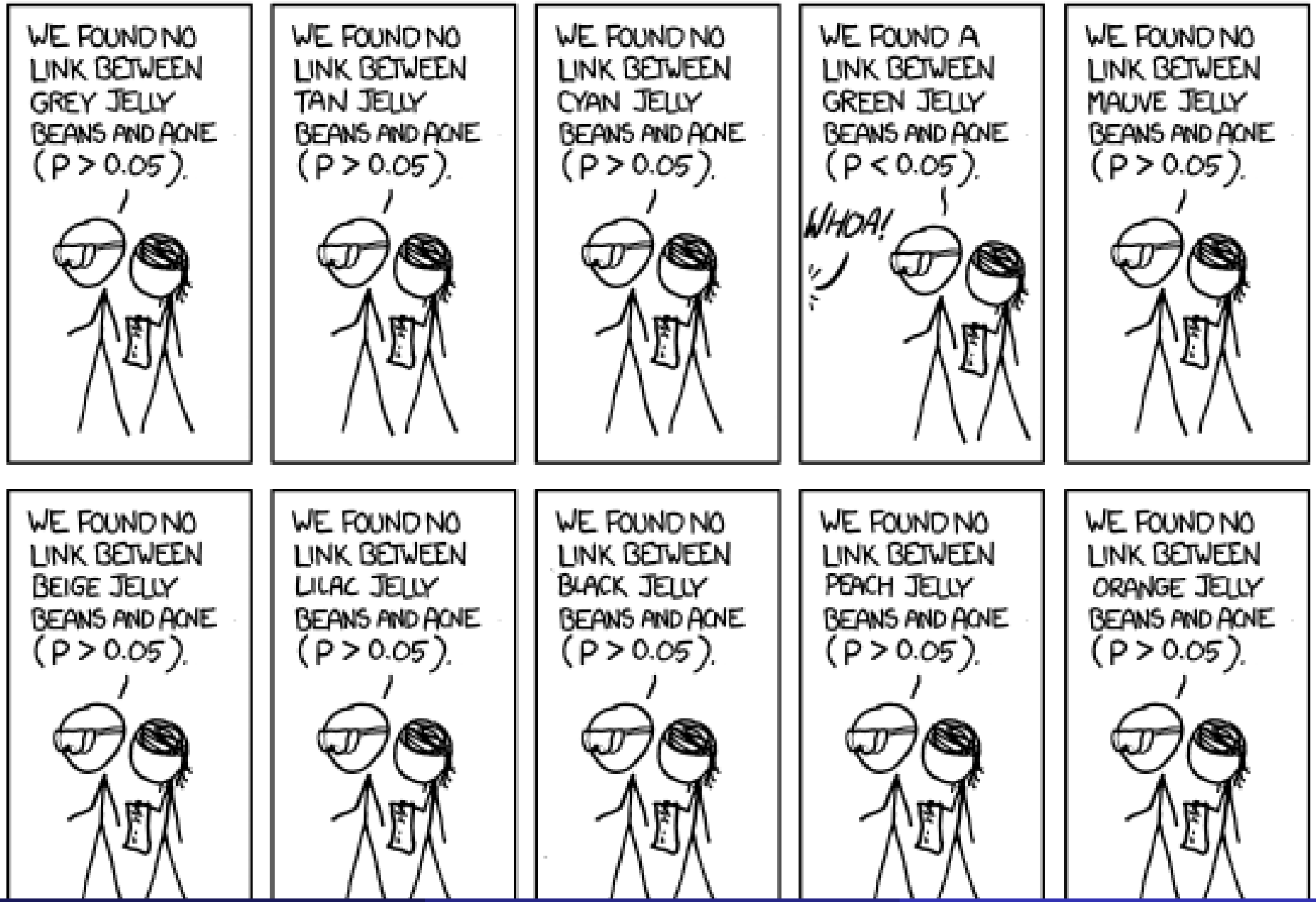
From XKCD (https://xkcd.com/882/)



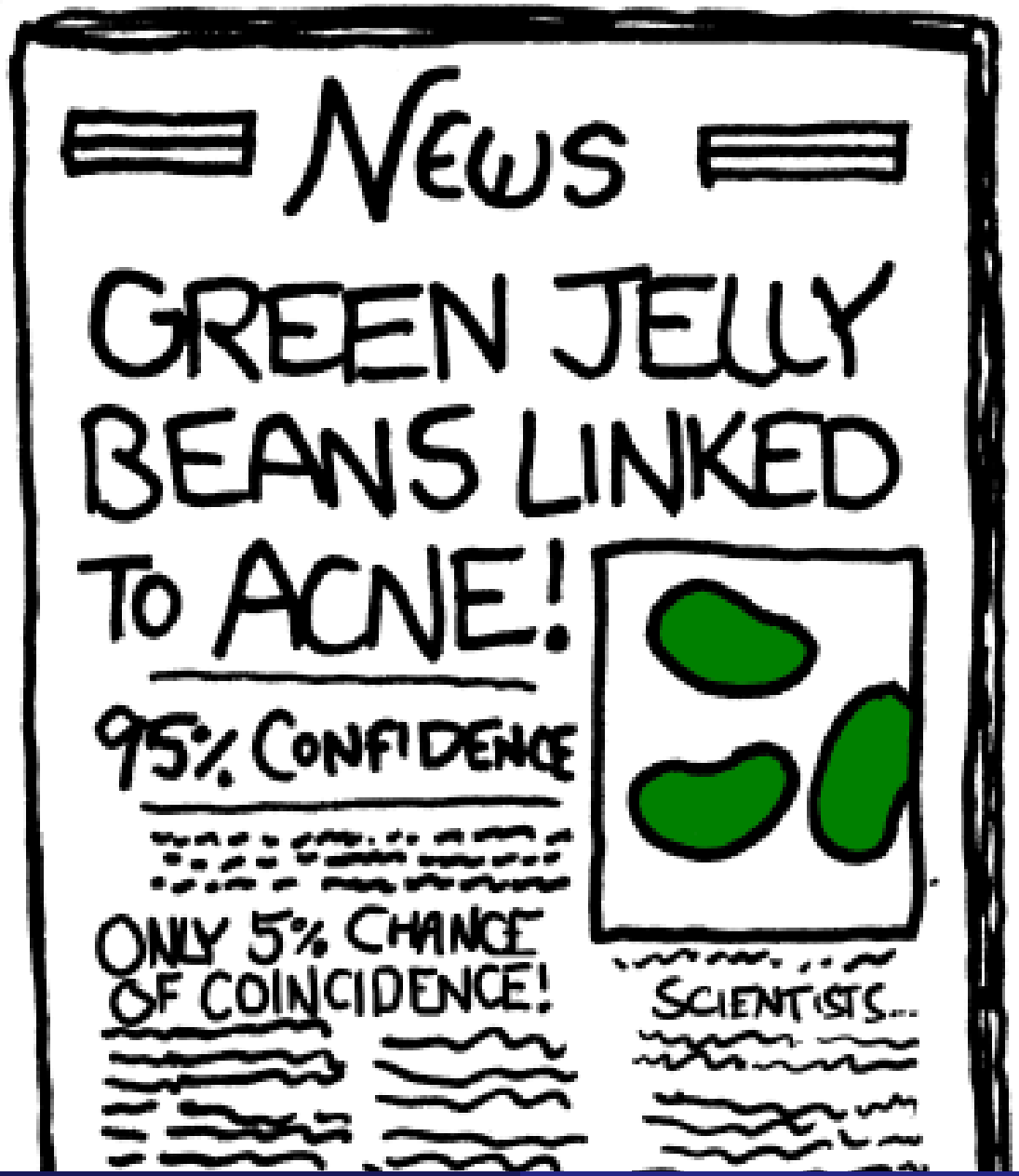
From XKCD (https://xkcd.com/882/)



From XKCD (https://xkcd.com/882/)



From XKCD (https://xkcd.com/882/)



From George Cobb - on why *p* values deserve to be re-evaluated

The **idea** of a p-value as one possible summary of evidence

morphed into a

- **rule** for authors: reject the null hypothesis if $p < .05$.

From George Cobb - on why *p* values deserve to be re-evaluated

The **idea** of a p-value as one possible summary of evidence

morphed into a

- **rule** for authors: reject the null hypothesis if $p < .05$,

which morphed into a

- **rule** for editors: reject the submitted article if $p > .05$.

From George Cobb - on why *p* values deserve to be re-evaluated

The **idea** of a p-value as one possible summary of evidence

morphed into a

- **rule** for authors: reject the null hypothesis if $p < .05$,

which morphed into a

- **rule** for editors: reject the submitted article if $p > .05$,

which morphed into a

- **rule** for journals: reject all articles that report p-values[1]

[1] http://www.nature.com/news/psychology-journal-bans-p-values-1.17001 describes the banning of null hypothesis significance testing by *Basic and Applied Psychology*.

From George Cobb - on why *p* values deserve to be re-evaluated

The **idea** of a p-value as one possible summary of evidence

morphed into a

- **rule** for authors: reject the null hypothesis if $p < .05$, which morphed into a
- **rule** for editors: reject the submitted article if $p > .05$, which morphed into a
- **rule** for journals: reject all articles that report p-values.

Bottom line: **Reject rules. Ideas matter.**

Posted to an American Statistical Association message board Oct 14 2015

Today's Agenda

Comparing Two Means Using Paired Samples

- Recognizing Paired (vs. Independent) Samples Designs
- Using `pivot_wider` or `pivot_longer` to reshape data
- Calculating Paired Differences
- Confidence Intervals (t, signed rank, bootstrap)
- Did Pairing Help Reduce Nuisance Variation?
- What Happens if we (incorrectly) use independent samples methods?

Today's Setup and Data

```
knitr::opts_chunk$set(comment = NA)
options(dplyr.summarise.inform = FALSE)

library(patchwork)
library(knitr)
library(magrittr)
library(janitor)
library(broom)
library(tidyverse)

theme_set(theme_bw())

dm431 <- readRDS("data/dm431_2020.Rds")
source("data/Love-boost.R")
```

Analyzing Changes in LDL Cholesterol

Comparing Means using Paired Samples

Our population: ALL adults ages 31-70 seen for care this year and two years ago who live in Northeast Ohio with a diabetes diagnosis.

Our sample: 431 of those people, drawn in a way we hope is representative (but certainly isn't random).

Suppose we want to compare the mean `ldl` cholesterol level for a set of subjects this year to the mean `ldl` for the same subjects two years ago.

dm431 Example A.

Here are the available data on:

- ldl = LDL cholesterol level (mg/dl) now
- ldl_old = LDL cholesterol level (mg/dl) two years ago

```
dm431 %>% select(subject, ldl, ldl_old) %>% head(5)
```

```
# A tibble: 5 x 3
  subject   ldl ldl_old
  <chr>   <int>   <int>
1 S-001     126      71
2 S-002     172     182
3 S-003     105     127
4 S-004     127      NA
5 S-005     100      86
```

Each subject (with complete data) provides ldl and ldl_old.

Deal with missingness in dm431 Example A

We'll assume MCAR and do a complete-case analysis.

```
dm431_A <- dm431 %>% filter(complete.cases(ldl, ldl_old))

mosaic::favstats(~ ldl, data = dm431_A)
```

```
 min Q1 median  Q3 max     mean       sd   n missing
  34 71     90 112 227 96.00815 33.5834 368       0
```

```
mosaic::favstats(~ ldl_old, data = dm431_A)
```

```
 min Q1 median  Q3 max     mean       sd   n missing
  31 72     90 115 244 96.57609 34.5199 368       0
```

Are these samples paired/matched or independent?

- Deciding whether or not the samples are paired (matched) is something we do before we analyze the data.
- The best way to establish whether a study uses paired or independent samples is to look for the **link** between the two measurements that creates paired differences.
- The question we're going to ask ourselves is

Does it make sense to calculate paired differences?

- The most common setting is a pre-post design, where each subject is measured before and after some exposure or intervention.
- The link then is the subject (who provides data before and after, so that calculating each subject's improvement or change makes sense.)

Paired or Independent Samples in dm431_A?

We want to compare the mean ldl cholesterol level for a set of subjects this year to the mean ldl for the same subjects two years ago.

```
dm431_A %>% select(subject, ldl, ldl_old) %>% head(3)
```

```
# A tibble: 3 x 3
  subject   ldl ldl_old
  <chr>   <int>   <int>
1 S-001     126      71
2 S-002     172     182
3 S-003     105     127
```

- What is the outcome? What are the exposure groups we are comparing?
- Does this design create paired samples or independent samples?
  - Does it make sense to calculate paired differences?
  - What is the link between ldl and ldl_old?

Paired Samples: Calculate Paired Differences

We want to compare the mean `ldl` cholesterol level for a set of subjects this year to the mean `ldl` for the same subjects two years ago.

```
dm431_A <- dm431_A %>%
  mutate(ldl_change = ldl - ldl_old)

dm431_A %>%
  select(subject, ldl, ldl_old, ldl_change) %>%
  tail(3)
```

```
# A tibble: 3 x 4
  subject   ldl ldl_old ldl_change
  <chr>   <int>   <int>      <int>
1 S-429     166     104         62
2 S-430      34      36         -2
3 S-431      77      67         10
```

Formatting the Data (Wide vs. Long)

Wide format (most appropriate for paired/matched samples)

subject	treatment1	treatment2
A	140	150
B	135	145
C	128	119

Long format (most appropriate for independent samples)

subject	sbp	group
A	140	treatment1
A	150	treatment2
B	135	treatment1
B	145	treatment2
C	128	treatment1
C	119	treatment2

Suppose you have a wide data set. . .

```
tempdat_wide <- tibble(
  subject = c("A", "B", "C"),
  treatment_1 = c(140, 135, 128),
  treatment_2 = c(150, 145, 119)
)

tempdat_wide
```

```
# A tibble: 3 x 3
  subject treatment_1 treatment_2
  <chr>         <dbl>       <dbl>
1 A               140         150
2 B               135         145
3 C               128         119
```

Pivot Data to make it longer

We want more rows, fewer columns. Each subject*treatment combination will become a row.

```
tempdat_long <- tempdat_wide %>%
  pivot_longer( -subject,
    names_to = "group", values_to = "sbp")
tempdat_long
```

```
# A tibble: 6 x 3
  subject group         sbp
  <chr>   <chr>       <dbl>
1 A       treatment_1   140
2 A       treatment_2   150
3 B       treatment_1   135
4 B       treatment_2   145
5 C       treatment_1   128
6 C       treatment_2   119
```

Pivot Data to make it wider

```
tempdat_wide2 <- tempdat_long %>%
  pivot_wider(names_from = group, values_from = sbp)

tempdat_wide2
```

```
# A tibble: 3 x 3
  subject treatment_1 treatment_2
  <chr>         <dbl>       <dbl>
1 A               140         150
2 B               135         145
3 C               128         119
```

Paired vs. Independent samples design?

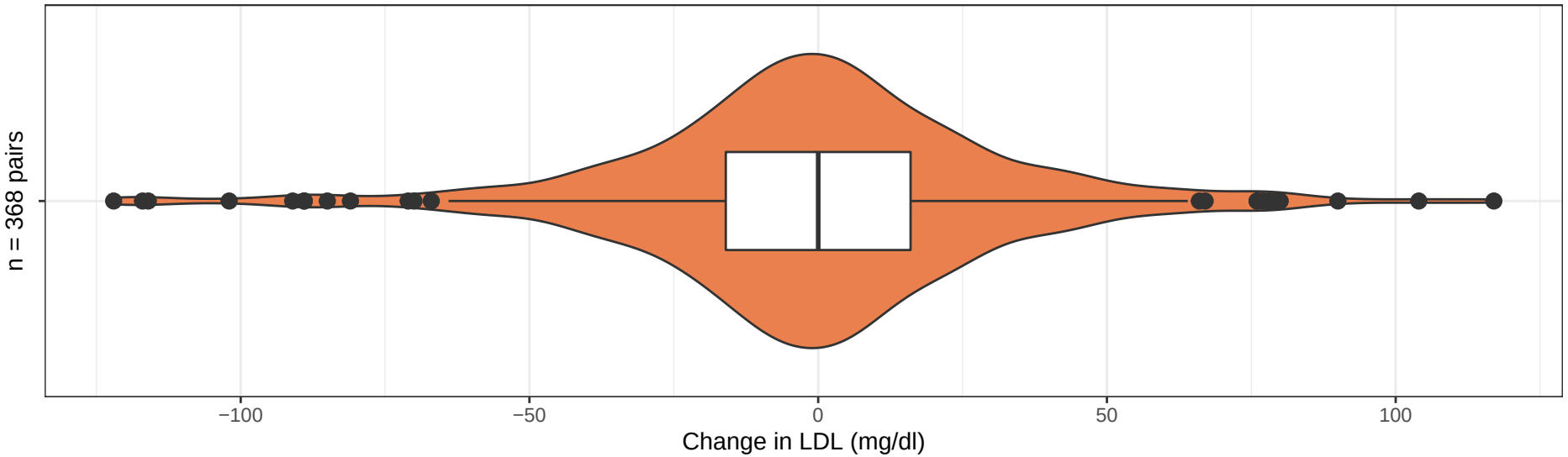
Deciding whether two samples are paired or independent is determined based solely on how the data are collected.

- Paired (matched) samples impose a matching, based on a link between the responses in one exposure group to the responses in the other exposure group.
  - The link is often the subject, measured under two different conditions, so that a subject-specific change is of interest.
  - The link can also be through some other sort of matching, where we match up two subjects, and then assign group 1 to one member of the pair and group 2 to the other member of the pair, so that each pair can be compared directly.
- Independent samples designs do not impose a matching, but instead sample two unrelated sets of subjects, where each group receives one of the two exposures.
- Paired (matched) samples designs require balanced samples (every measurement must be part of one and only one pair) while independent samples do not.

Returning to our `dm431_A` comparison

We now have a sample of paired differences (LDL now - LDL two years ago). Here are some summaries:

min	Q1	median	Q3	max	mean	sd	n	missing
-122	-16	0	16	117	-0.57	32.24	368	0



Building Confidence Intervals for Paired Samples

is identical to building confidence intervals for a single population mean.

1. A **t-based** estimate and confidence interval, available from an intercept-only linear model, or (equivalently) a t test.
    - This approach will require an assumption that the population comes from a Normal distribution.
2. A **bootstrap** confidence interval, which uses resampling to estimate the population mean.
    - This approach won't require the Normality assumption, but has some other constraints.
3. A **Wilcoxon signed rank** approach, but that won't describe the mean, only a pseudo-median.
    - This also doesn't require the Normality assumption, but no longer describes the population mean (or median) unless the population can be assumed symmetric. Instead it describes the *pseudo-median*.

It's just the one-sample situation again, but with paired differences.

Intercept-only Regression for the Paired Differences

We'll build a 90% confidence interval for the population mean change in LDL using the t distribution with an indicator variable regression. It's just a linear model.

```
model_A <- lm(ldl_change ~ 1, data = dm431_A)
tidy(model_A, conf.int = TRUE, conf.level = 0.9) %>%
    select(term, estimate, conf.low, conf.high, p.value)
```

```
# A tibble: 1 x 5
  term        estimate conf.low conf.high p.value
  <chr>          <dbl>    <dbl>     <dbl>   <dbl>
1 (Intercept)   -0.568    -3.34      2.20   0.736
```

Could also do this with . . .

```
dm431_A %$%
  t.test(ldl, ldl_old, paired = TRUE, conf.level = 0.9)
```

```
    Paired t-test

data:  ldl and ldl_old
t = -0.33795, df = 367, p-value = 0.7356
alternative hypothesis: true difference in means is not equal
90 percent confidence interval:
 -3.339140  2.203271
sample estimates:
mean of the differences
             -0.5679348
```

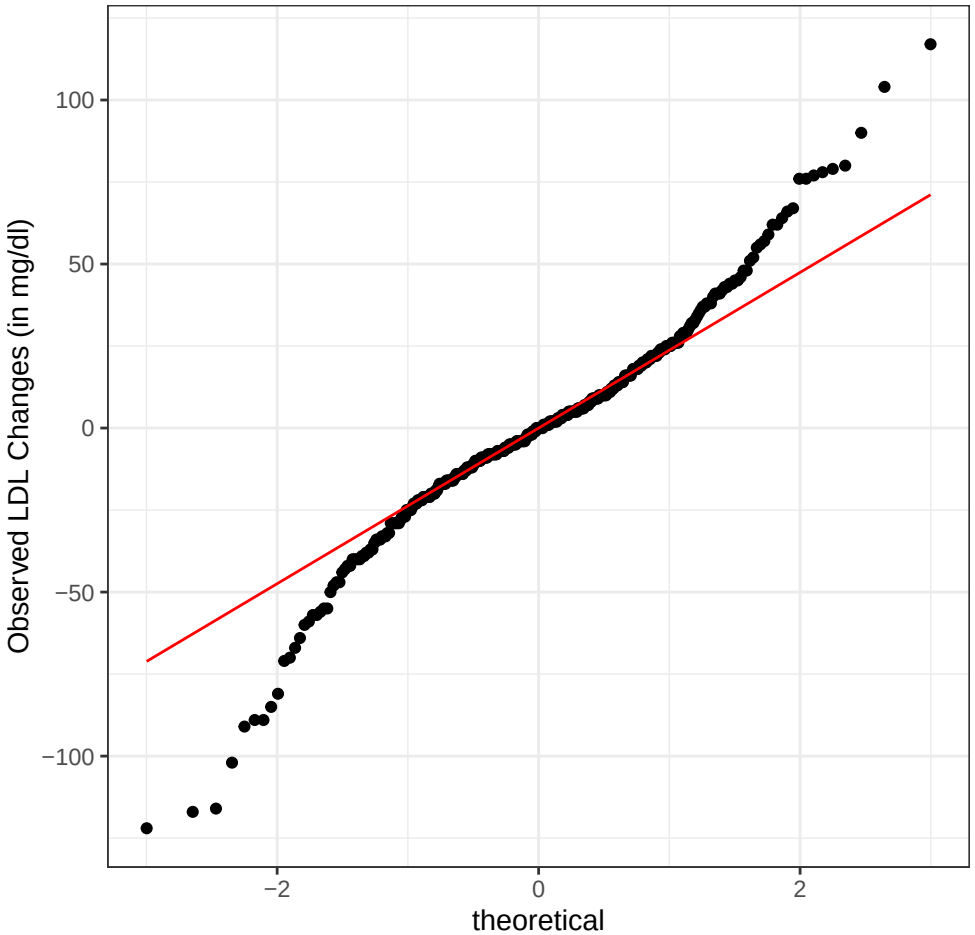
Or we could do this with. . .

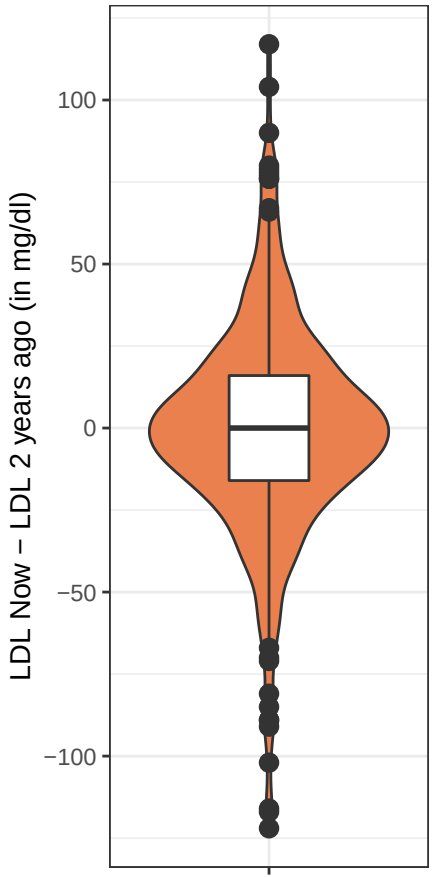
```
tt <- dm431_A %$%
  t.test(ldl_change, conf.level = 0.9) %>% tidy()

tt %>%
  select(method, alternative, estimate, conf.low, conf.high)
```

```
# A tibble: 1 x 5
  method          alternative estimate conf.low conf.high
  <chr>           <chr>          <dbl>    <dbl>     <dbl>
1 One Sample t~ two.sided       -0.568    -3.34      2.20
```

Can we assume Normality in LDL changes?

LDL Changes (Now − 2 Years Ago) in dm431_A





Wilcoxon Signed Rank procedure

Supposing we don't want to assume Normality, but are willing to assume symmetry.

```
dm431_A %$%
  wilcox.test(ldl - ldl_old, conf.int=TRUE, conf.level = 0.9)
```

```
    Wilcoxon signed rank test with continuity
    correction

data:  ldl - ldl_old
V = 31966, p-value = 0.791
alternative hypothesis: true location is not equal to 0
90 percent confidence interval:
 -2.500002  1.999938
sample estimates:
(pseudo)median
    -0.4999817
```

Bootstrap CI for the Changes in LDL

What if we're not willing to assume Normality or symmetry, but still want to compare means?

```
set.seed(20201027)
Hmisc::smean.cl.boot(dm431_A$ldl_change, conf.int = 0.9)
```

```
      Mean       Lower       Upper
-0.5679348 -3.4144022  2.2410326
```

What does this confidence interval suggest about the p value?

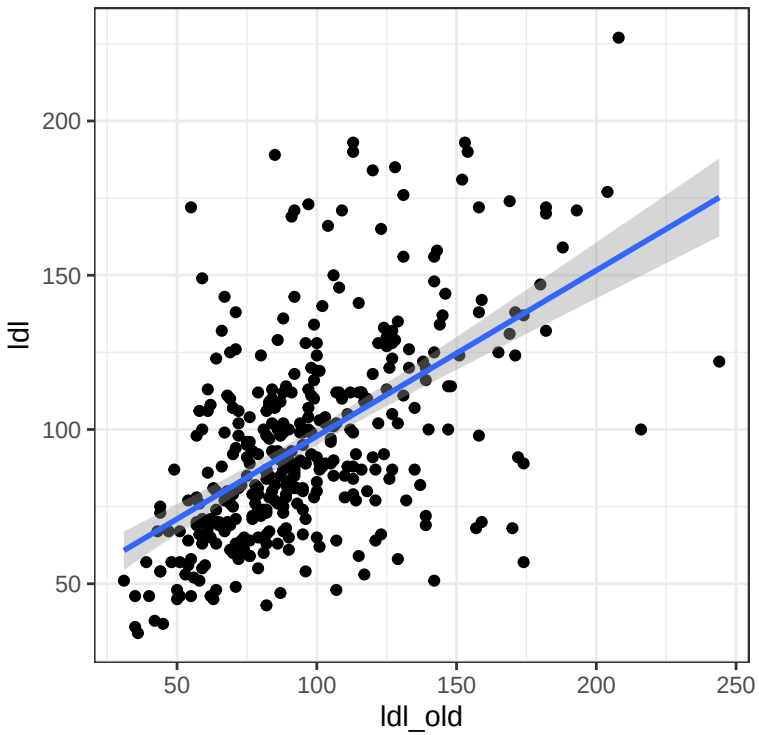
Paired Samples: Findings from `dm431_A`

Define μ_d = population mean difference (LDL now minus LDL two years ago)

Approach	p value	95% CI for μ_d
t Test	0.7356	(-3.3, 2.2)
Wilcoxon	0.7910	(-2.5, 2.0)
Bootstrap	> 0.10	(-3.4, 2.2)

Are the pairs of measurements positively associated?

```
ggplot(dm431_A, aes(x = ldl_old, y = ldl)) +
  geom_point() +
  geom_smooth(method = "lm", formula = y ~ x) +
  theme(aspect.ratio = 1) # for slide presentation
```



Did pairing help in this situation to reduce noise?

If the correlation of `ldl` and `ldl_old` is substantial and positive, then pairing helps account for this nuisance variation.

```
dm431_A %$% cor(ldl, ldl_old)
```

```
[1] 0.5521376
```

Was there a positive correlation of `ldl` and `ldl_old`?

- Yes, it was 0.55, so there was some reduction in nuisance variation at the subject level.

What if we did this (incorrectly) assuming independent samples?

We would need to rearrange the data to let us look at the samples as if they were independent. Specifically, we would have to pivot to create a longer data set.

```
dm_ldl_longer <-
  dm431_A %>% select(subject, ldl, ldl_old) %>%
  pivot_longer(
    cols = starts_with("ldl"),
    names_to = "time", values_to = "LDL")
```

The dm_ldl_longer data

Let's just look at four rows of the full data set.

```
dm_ldl_longer %>% filter(subject %in% c("S-029", "S-131"))
```

```
# A tibble: 4 x 3
  subject time       LDL
  <chr>   <chr>    <int>
1 S-029   ldl         48
2 S-029   ldl_old     64
3 S-131   ldl         55
4 S-131   ldl_old     79
```

Summarizing the data as independent samples

```
mosaic::favstats(LDL ~ time, data = dm_ldl_longer) %>%
  kable(dig = 2)
```

time	min	Q1	median	Q3	max	mean	sd	n	missing
ldl	34	71	90	112	227	96.01	33.58	368	0
ldl_old	31	72	90	115	244	96.58	34.52	368	0

Using dm_ldl_longer: independent samples t test

```
model2 <- lm(LDL ~ time, data = dm_ldl_longer)
tidy(model2, conf.int = TRUE, conf.level = 0.90) %>%
  select(term, estimate, conf.low, conf.high, p.value)
```

```
# A tibble: 2 x 5
  term         estimate conf.low conf.high   p.value
  <chr>           <dbl>    <dbl>     <dbl>     <dbl>
1 (Intercept)   96.0       93.1      98.9  3.02e-258
2 timeldl_old    0.568     -3.57      4.70 8.21e-  1
```

Inappropriate Results (treat samples as independent)

Comparing the LDL for the current data (now) to the previous data (old) without accounting for the fact that the same people provided the data in each sample.

Procedure	p for $H_0 : \mu_{now} = \mu_{old}$	90% CI for $\mu_{now} - \mu_{old}$
Pooled t test	0.82	(-4.7, 3.6)
Welch t test	0.82	(-4.7, 3.6)
Rank Sum test	0.86	(-4.0, 3.0)
Bootstrap CI	$p > 0.10$	(-4.7, 3.7)

- What changes here when we (incorrectly) ignore the pairing?

Note I used the seed 2020 to obtain the bootstrap result.

A second study

Suppose we look at `dbp` (diastolic blood pressure) instead of `ldl` in this setting, and compare `dbp` now to `dbp` 2 years ago.

- Again, these are paired samples.
- A difference is that we have no missing data in `dbp` or `dbp_old`.

Summary of results:

Paired Samples Study Designs

- Using a paired samples design means we carefully sample matched sets of subjects in pairs, so that the sampled subjects in each pair are as similar as possible, except for the exposure of interest.
- Each observation in one exposure group is matched to a single observation in the other exposure group, so that taking paired differences is a rational thing to do.
- Since every subject must be matched to exactly one subject in the other group, the sizes of the groups must be equal.
- If the data are collected using paired samples, we should use a paired samples analysis.

On "Significance"

- **A significant effect is not the same thing as an interesting effect.** For example, results calculated from large samples are nearly always "significant" even when the effects are quite small in magnitude. Before doing a test, always ask if the effect is large enough to be of any practical interest. If not, why do the test?
- **A non-significant effect is not the same thing as no difference.** A large effect of real practical interest may still produce a non-significant result simply because the sample is too small.
- **There are assumptions behind all statistical inferences.** Checking assumptions is crucial to validating the inference made by any test or confidence interval.

Next Time

Comparing Population Rates/Proportions/Percentages

For Self-Study: Analyzing Changes in Diastolic BP

dm431 Example B. (Diastolic BP changes)

Here are the available data on:

- dbp = Diastolic Blood Pressure (mm Hg) now
- dbp_old = Diastolic Blood Pressure (mm Hg) two years ago

```
dm431 %>% select(subject, dbp, dbp_old) %>% head(5)
```

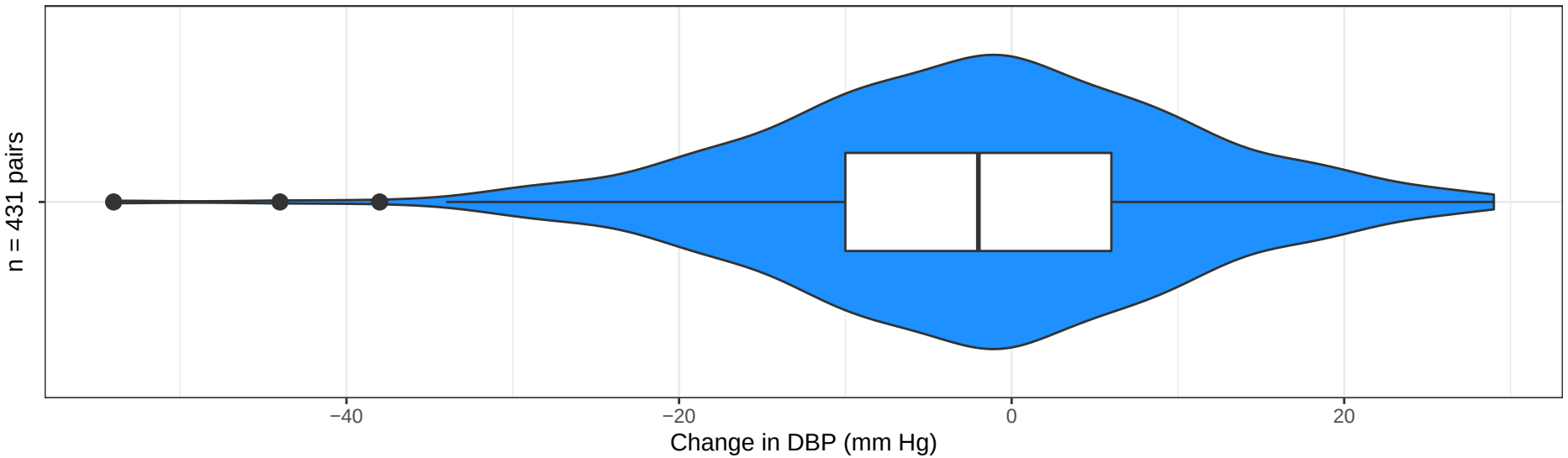
```
# A tibble: 5 x 3
  subject   dbp dbp_old
  <chr>   <int>   <int>
1 S-001      64      70
2 S-002      84      60
3 S-003      95      92
4 S-004      87      70
5 S-005      58      72
```

Each subject has complete data on dbp and dbp_old.

Paired Samples? Calculate Paired Differences

```
dm431 <- dm431 %>%
    mutate(dbp_chg = dbp - dbp_old)

mosaic::favstats(~ dbp_chg, data = dm431) %>% round(., 2)
```

```
 min  Q1 median Q3 max  mean    sd   n missing
 -54 -10     -2  6  29 -2.03 12.42 431       0
```



Intercept-only Regression for the Paired Differences

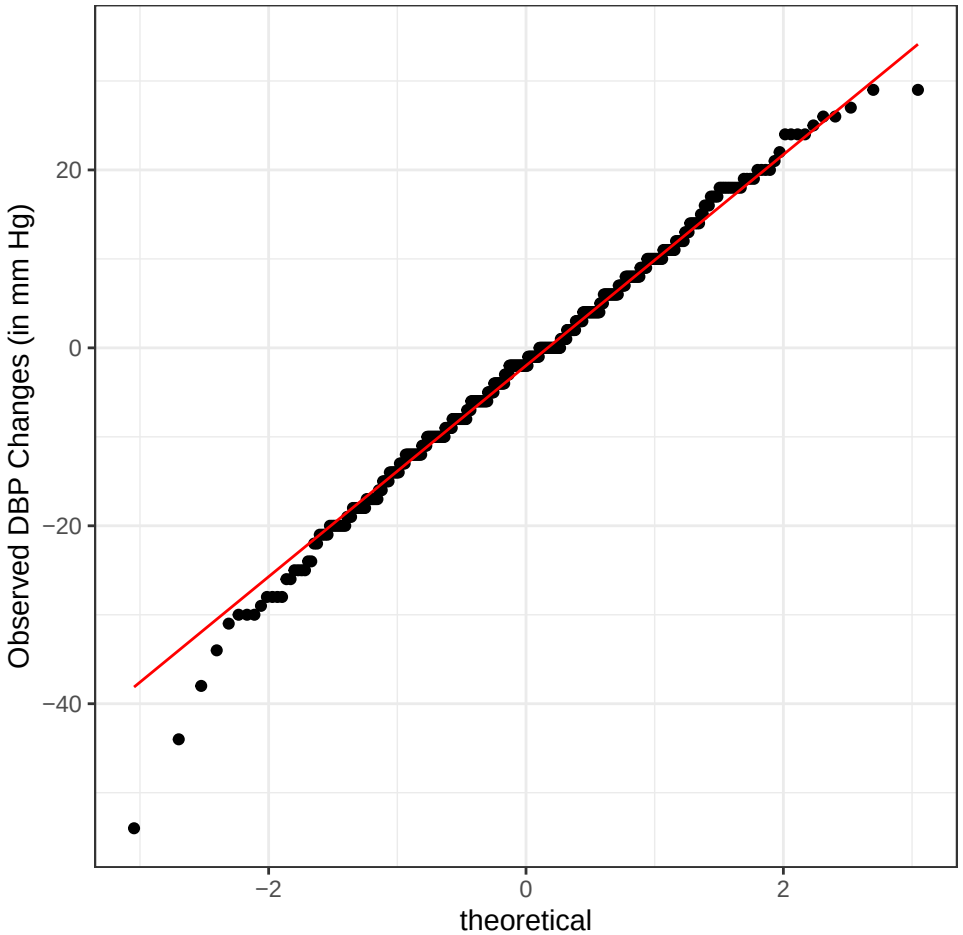
We'll build a 99% confidence interval for the population mean change (DBP now - DBP old), recalling that the point estimate (sample mean difference) was negative.

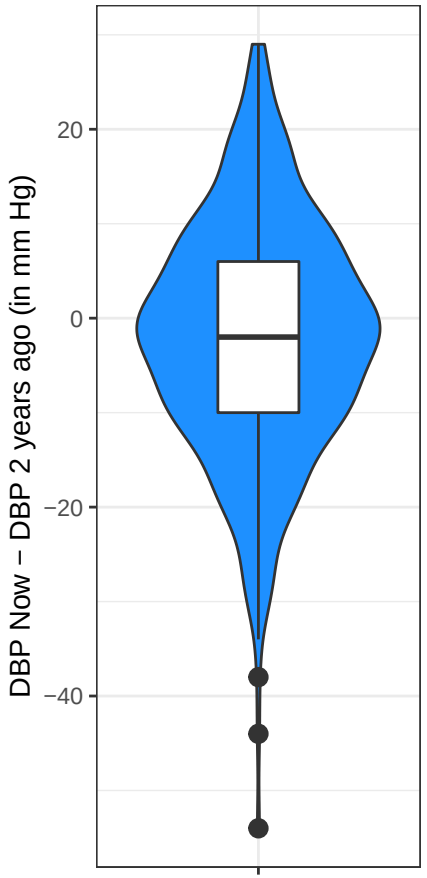
```
model_B <- lm(dbp_chg ~ 1, data = dm431)
tidy(model_B, conf.int = TRUE, conf.level = 0.99) %>%
    select(term, estimate, conf.low, conf.high, p.value)
```

```
# A tibble: 1 x 5
  term        estimate conf.low conf.high  p.value
  <chr>          <dbl>    <dbl>     <dbl>    <dbl>
1 (Intercept)    -2.03    -3.57    -0.478 0.000772
```

Can we assume Normality in DBP changes?

DBP Changes (Now − 2 Years Ago) in dm431





Wilcoxon Signed Rank procedure

Supposing we are willing to assume symmetry and treat a pseudo-median like a mean.

```
dm431 %$% wilcox.test(dbp_chg, conf.int=TRUE, conf.level = 0.9
```

```
    Wilcoxon signed rank test with continuity
    correction

data:  dbp_chg
V = 33462, p-value = 0.001962
alternative hypothesis: true location is not equal to 0
99 percent confidence interval:
 -3.5000052 -0.4999952
sample estimates:
(pseudo)median
     -2.000001
```

Bootstrap CI for the Changes in LDL

What if we're not willing to assume Normality or symmetry, but still want to compare means?

```
set.seed(20201027)
Hmisc::smean.cl.boot(dm431$dbp_chg, conf.int = 0.99)
```

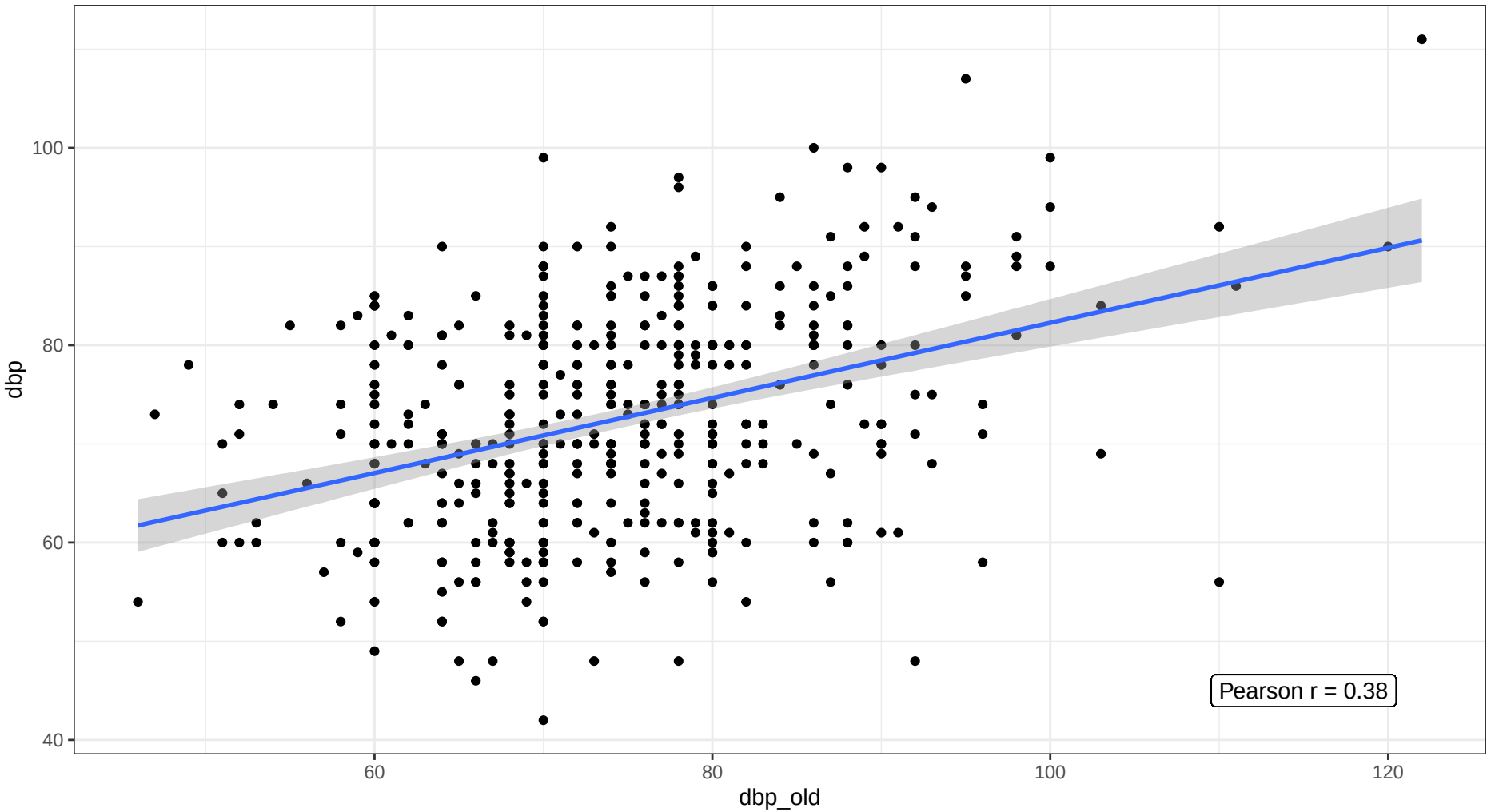
```
      Mean       Lower       Upper
-2.0255220 -3.6496868 -0.4894432
```

Paired Samples: Findings from dm431

Define μ_d = population mean difference (DBP now minus DBP two years ago)

Approach	p value	95% CI for μ_d
t Test	0.0008	(-3.57, -0.48)
Wilcoxon	0.0020	(-3.50, -0.50)
Bootstrap	< 0.01	(-3.65, -0.49)

Are the pairs of measurements positively associated?



Code for previous slide

```
ggplot(dm431, aes(x = dbp_old, y = dbp)) +
  geom_point() +
  geom_smooth(method = "lm", formula = y ~ x) +
  geom_label(x = 115, y = 45,
             label =
               paste0("Pearson r = ",
                      dm431 %$% cor(dbp, dbp_old) %>%
                        round_half_up(., 2)))
```

Did pairing help in this situation to reduce noise?

If the correlation of dbp and dbp_old is substantial and positive, then pairing helps account for this nuisance variation.

```
dm431 %$% cor(dbp, dbp_old)
```

```
[1] 0.3838496
```

Was there a positive correlation of dbp and dbp_old?

- Yes, it was 0.38, so there was some reduction in nuisance variation at the subject level.