A sense of proportion

HYPOTHESIS TESTING IN R

Richie Cotton
Data Evangelist at DataCamp

Chapter 1 recap

- Is a claim about an unknown population proportion feasible?
- Standard error of sample statistic calculated using bootstrap distribution.
- This was used to compute a standardized test statistic, ...
- which was used to calculate a p-value, ...
- which was used to decide which hypothesis made most sense.
- Here, we'll calculate the test statistic without using the bootstrap distribution.

Standardized test statistic for proportions

p: population proportion (unknown population parameter)

$\hat{p}$: sample proportion (sample statistic)

p_0: hypothesized population proportion

$$z = \frac{\hat{p} - \text{mean}(\hat{p})}{\text{standard error}(\hat{p})} = \frac{\hat{p} - p}{\text{standard error}(\hat{p})}$$

Assuming H_0 is true, $p = p_0$, so

$$z = \frac{\hat{p} - p_0}{\text{standard error}(\hat{p})}$$

Easier standard error calculations

$$SE(\bar{x}_{\text{child}} - \bar{x}_{\text{adult}}) \approx \sqrt{\frac{s^2_{\text{child}}}{n_{\text{child}}} + \frac{s^2_{\text{adult}}}{n_{\text{adult}}}}$$

$$SE_{\hat{p}} = \sqrt{\frac{p_0 * (1 - p_0)}{n}}$$

Assuming H_0 is true,

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 * (1 - p_0)}{n}}}$$

This only uses sample information ($\hat{p}$ and n) and the hypothesized parameter (p_0).

Why z instead of t?

$$t = \frac{(\bar{x}_{\text{child}} - \bar{x}_{\text{adult}})}{\sqrt{\frac{s_{\text{child}}^2}{n_{\text{child}}} + \frac{s_{\text{adult}}^2}{n_{\text{adult}}}}}$$

- s is calculated from $\bar{x}$, so $\bar{x}$ is used to estimate the population mean *and* to estimate the population standard deviation.
- This increases uncertainty in our estimate of the population parameter.
- t-distribution has fatter tails than a normal distribution.
- This gives an extra level of caution.
- $\hat{p}$ only appears in the numerator, so z-scores are fine.

Stack Overflow age categories

H_0: The proportion of SO users under thirty is equal to 0.5.

H_A: The proportion of SO users under thirty is not equal to 0.5.

```
alpha <- 0.01
```

```
stack_overflow %>%
  count(age_cat)
```

```
# A tibble: 2 x 2
  age_cat          n
  <chr>        <int>
1 At least 30   1050
2 Under 30      1216
```

Variables for z

```
p_hat <- stack_overflow %>%
  summarize(prop_under_30 = mean(age_cat == "Under 30")) %>%
  pull(prop_under_30)
```

```
0.5366
```

```
p_0 <- 0.50
```

```
n <- nrow(stack_overflow)
```

```
2266
```

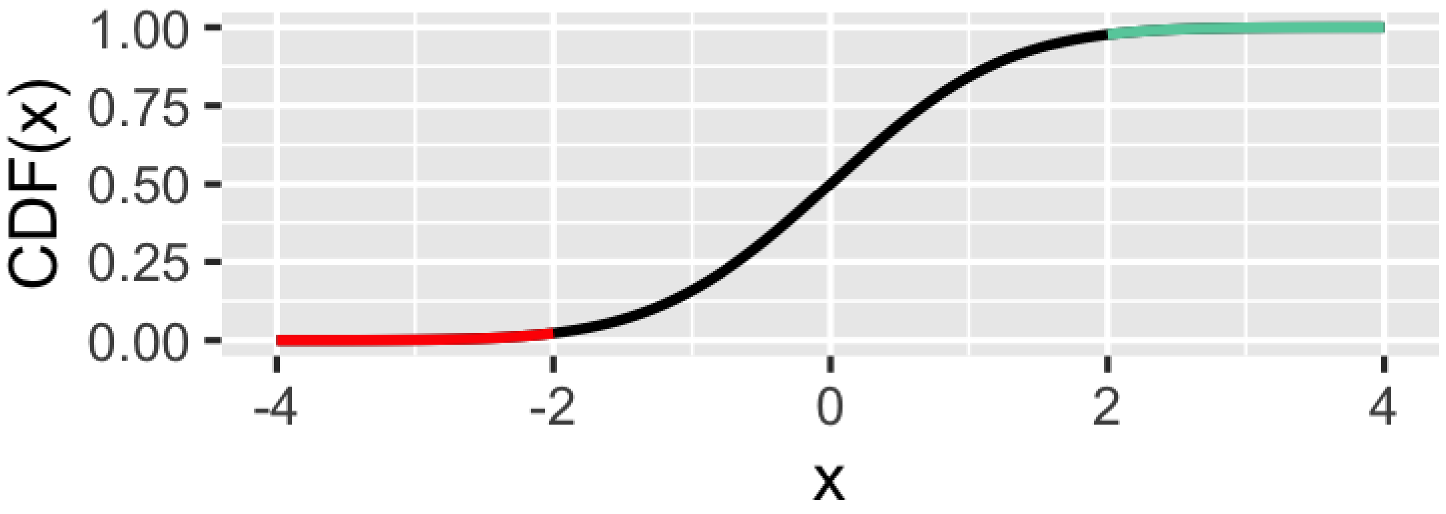
Calculating the z-score

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 * (1 - p_0)}{n}}}$$

```
numerator <- p_hat - p_0
denominator <- sqrt(p_0 * (1 - p_0) / n)
z_score <- numerator / denominator
```

```
3.487
```

Calculating the p-value



Left-tailed ("less than")

```
p_value <- pnorm(z_score)
```

Right-tailed ("greater than")

```
p_value <- pnorm(z_score, lower.tail = FALSE)
```

Two-tailed ("not equal")

```
p_value <- pnorm(z_score) +
  pnorm(z_score, lower.tail = FALSE)
```

```
p_value <- 2 * pnorm(z_score)
```

```
0.000244
```

```
p_value <= alpha
```

```
TRUE
```

Let's practice!

HYPOTHESIS TESTING IN R

A sense of proportion

HYPOTHESIS TESTING IN R

Richie Cotton
Data Evangelist at DataCamp

Comparing two proportions

H_0: The proportion of SO users who are hobbyists is the same for those under thirty as those at least thirty.

$$H_0: p_{\geq 30} - p_{<30} = 0$$

H_A: The proportion of SO users who are hobbyists is different for those under thirty as those at least thirty.

$$H_A: p_{\geq 30} - p_{<30} \neq 0$$

```
alpha <- 0.05
```

Calculating the z-score

$$z = \frac{(\hat{p}_{\geq 30} - \hat{p}_{<30}) - 0}{\text{SE}(\hat{p}_{\geq 30} - \hat{p}_{<30})}$$

$$\text{SE}(\hat{p}_{\geq 30} - \hat{p}_{<30}) = \sqrt{\frac{\hat{p} \times (1 - \hat{p})}{n_{\geq 30}} + \frac{\hat{p} \times (1 - \hat{p})}{n_{<30}}}$$

$\hat{p}$ is a *pooled estimate* for p (common unknown proportion of successes).

$$\hat{p} = \frac{n_{\geq 30} \times \hat{p}_{\geq 30} + n_{<30} \times \hat{p}_{<30}}{n_{\geq 30} + n_{<30}}$$

We only need to calculate 4 numbers: $\hat{p}_{\geq 30}$, $\hat{p}_{<30}$, $n_{\geq 30}$, $n_{<30}$.

Getting the numbers for the z-score

```
stack_overflow %>%
  group_by(age_cat) %>%
  summarize(
    p_hat = mean(hobbyist == "Yes"),
    n = n()
  )
```

```
# A tibble: 2 x 3
  age_cat      p_hat     n
  <chr>        <dbl> <int>
1 At least 30 0.773  1050
2 Under 30    0.843  1216
```

```
z_score
```

```
-4.217
```

Proportion tests using prop_test()

```
library(infer)
stack_overflow %>%
  prop_test(
    hobbyist ~ age_cat,                      # proportions ~ categories
    order = c("At least 30", "Under 30"), # which p-hat to subtract
    success = "Yes",                         # which response value to count proportions of
    alternative = "two-sided",               # type of alternative hypothesis
    correct = FALSE                          # should Yates' continuity correction be applied?
  )
```

```
# A tibble: 1 x 6
  statistic chisq_df   p_value alternative lower_ci upper_ci
      <dbl>    <dbl>     <dbl> <chr>          <dbl>    <dbl>
1      17.8        1 0.0000248 two.sided     0.0605    0.165
```

Let's practice!

HYPOTHESIS TESTING IN R

Declaration of independence

HYPOTHESIS TESTING IN R

Richie Cotton
Data Evangelist at DataCamp

Revisiting the proportion test

```
library(infer)
stack_overflow %>%
  prop_test(
    hobbyist ~ age_cat,
    order = c("At least 30", "Under 30"),
    alternative = "two-sided",
    correct = FALSE
  )
```

```
# A tibble: 1 x 6
  statistic chisq_df   p_value alternative lower_ci upper_ci
      <dbl>    <dbl>     <dbl> <chr>          <dbl>    <dbl>
1      17.8        1 0.0000248 two.sided     0.0605    0.165
```

Independence of variables

Previous hypothesis test result: there is evidence that the `hobbyist` and `age_cat` variables have an association.

If the proportion of successes in the response variable is the same across all categories of the explanatory variable, the two variables are *statistically independent*.

[1] Response and explanatory variables are defined in "Introduction to Regression in R", Chapter 1.

Job satisfaction and age category

```
stack_overflow %>%
  count(age_cat)
```

```
# A tibble: 2 x 2
  age_cat         n
  <chr>       <int>
1 At least 30  1050
2 Under 30     1211
```

```
stack_overflow %>%
  count(job_sat)
```

```
# A tibble: 5 x 2
  job_sat                   n
  <fct>                 <int>
1 Very dissatisfied       159
2 Slightly dissatisfied   342
3 Neither                 201
4 Slightly satisfied      680
5 Very satisfied          879
```

Declaring the hypotheses

H_0: Age categories are independent of job satisfaction levels.

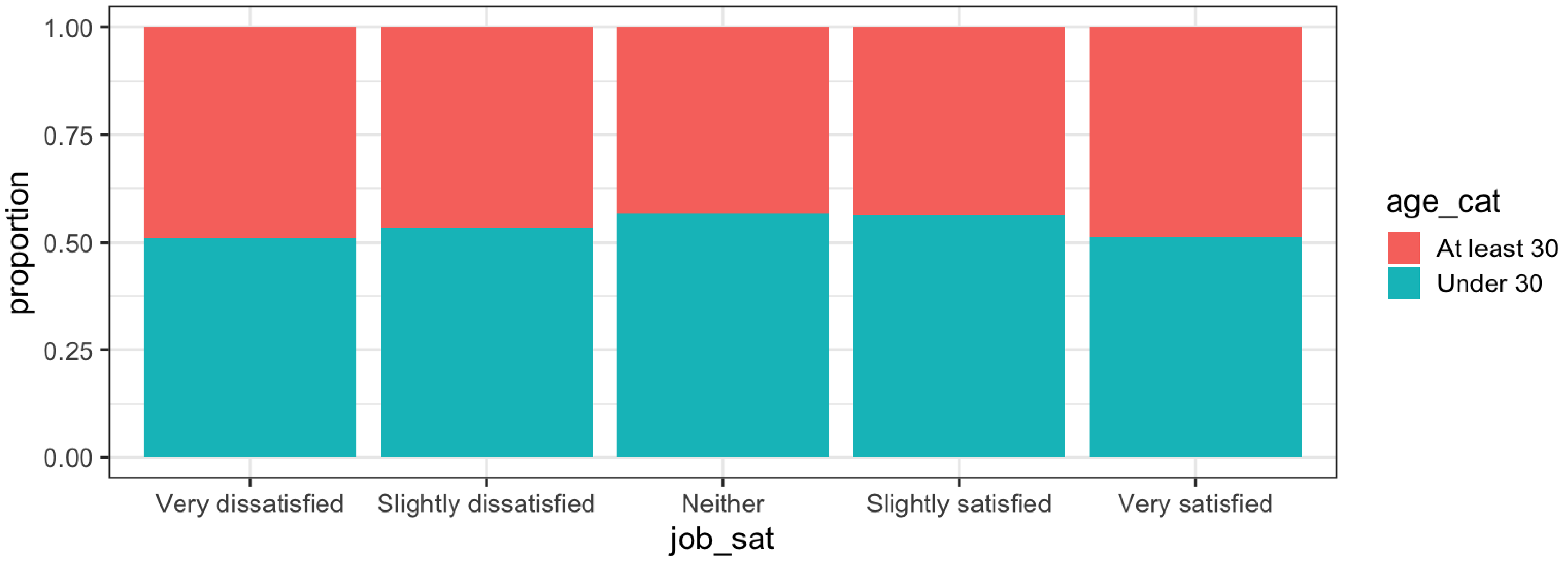
H_A: Age categories are not independent of job satisfaction levels.

```
alpha <- 0.1
```

- Test statistic denoted χ^2.
- Assuming independence, how far away are the observed results from the expected values?

Exploratory visualization: proportional stacked bar plot

```
ggplot(stack_overflow, aes(job_sat, fill = age_cat)) +
  geom_bar(position = "fill") +
  ylab("proportion")
```



Chi-square independence test using chisq_test()

```
library(infer)
stack_overflow %>%
  chisq_test(age_cat ~ job_sat)
```

```
# A tibble: 1 x 3
  statistic chisq_df p_value
      <dbl>    <int>   <dbl>
1      5.55        4   0.235
```

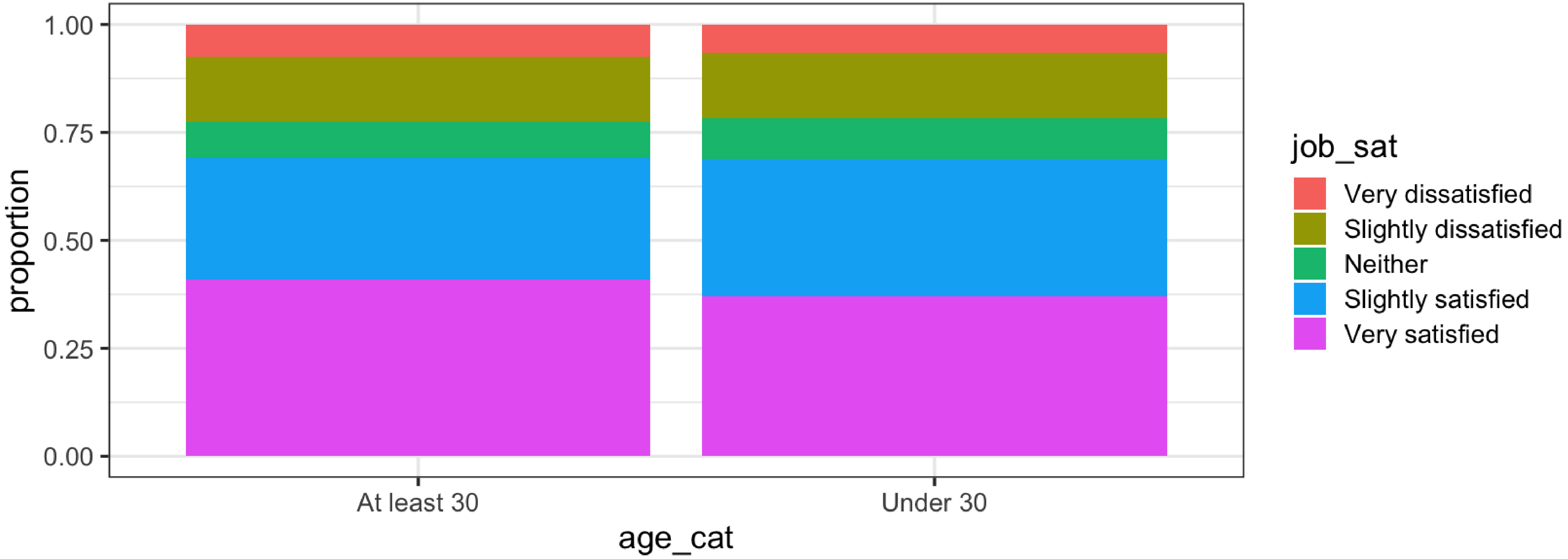
Degrees of freedom:

$$(\text{No. of response categories} - 1) \times (\text{No. of explanatory categories} - 1)$$

$$(2 - 1) * (5 - 1) = 4$$

Swapping the variables?

```
ggplot(stack_overflow, aes(age_cat, fill = job_sat)) +
  geom_bar(position = "fill") +
  ylab("proportion")
```



chi-square both ways

```
library(infer)
stack_overflow %>%
  chisq_test(age_cat ~ job_sat)
```

```
# A tibble: 1 x 3
  statistic chisq_df p_value
      <dbl>    <int>   <dbl>
1      5.55        4   0.235
```

Ask

> Are the variables X and Y independent?

```
library(infer)
stack_overflow %>%
  chisq_test(job_sat ~ age_cat)
```

```
# A tibble: 1 x 3
  statistic chisq_df p_value
      <dbl>    <int>   <dbl>
1      5.55        4   0.235
```

Not

> Is variable X independent from variable Y?

What about direction and tails?

```
args(chisq_test)
```

```
function (x, formula, response = NULL, explanatory = NULL, ...)
```

- Observed and expected counts squared must be non-negative.
- chi-square tests are almost always right-tailed. [1]

[1] Left-tailed chi-square tests are used in statistical forensics to detect is a fit is suspiciously good because the data was fabricated. Chi-square tests of variance can be two-tailed. These are niche uses though.

Let's practice!

HYPOTHESIS TESTING IN R

Does this dress make my fit look good?

HYPOTHESIS TESTING IN R

Richie Cotton
Data Evangelist at DataCamp

Purple links

> You search for a coding solution online and the first result link is purple because you already visited it. How do you feel?

```
purple_link_counts <- stack_overflow %>%
  count(purple_link)
```

```
# A tibble: 4 x 2
  purple_link           n
  <fct>             <int>
1 Hello, old friend  1330
2 Amused              409
3 Indifferent         426
4 Annoyed             290
```

Declaring the hypotheses

```r
hypothesized <- tribble(
  ~ purple_link, ~ prop,
  "Hello, old friend", 1 / 2,
  "Amused"           , 1 / 6,
  "Indifferent"      , 1 / 6,
  "Annoyed"          , 1 / 6
)
```

```
# A tibble: 4 x 2
  purple_link         prop
  <chr>              <dbl>
1 Hello, old friend 0.5
2 Amused            0.167
3 Indifferent       0.167
4 Annoyed           0.167
```

H_0: The sample matches with the hypothesized distribution.

H_A: The sample does not match with the hypothesized distribution.

The test statistic, χ^2, measures how far observed results are from expectations in each group.

```r
alpha <- 0.01
```

[1] tribble is short for "row-wise tibble"; not to be confused with the alien species from Star Trek

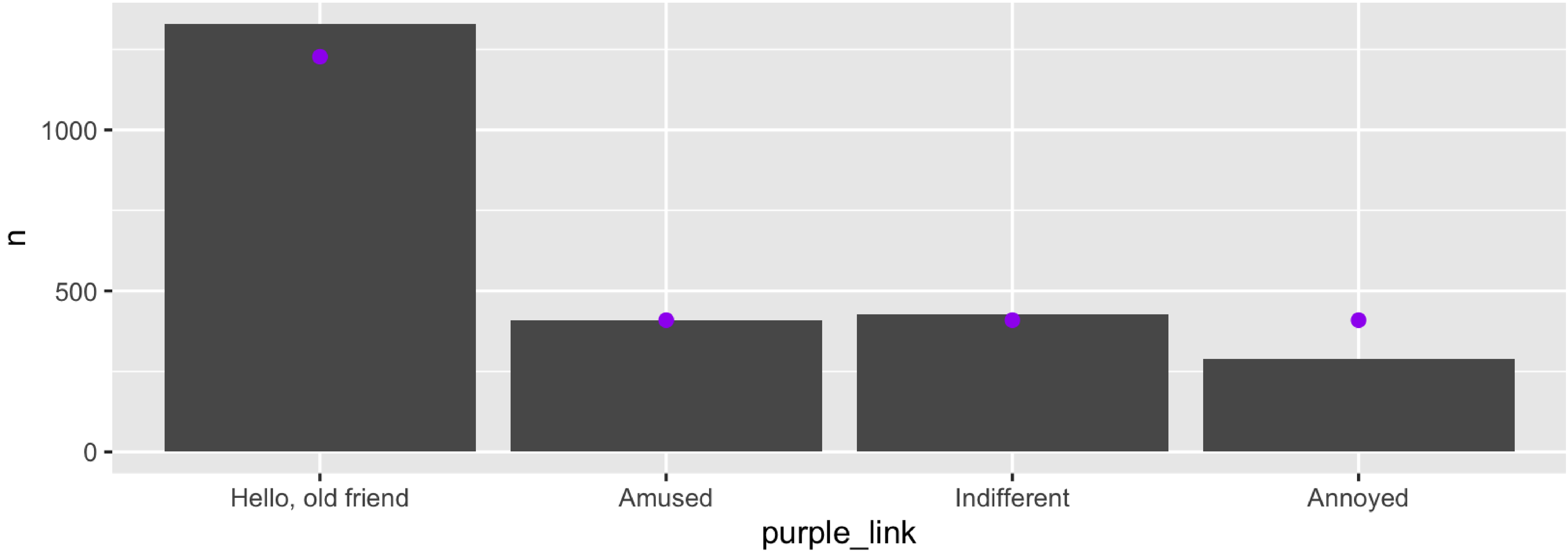
Hypothesized counts by category

```
n_total <- nrow(stack_overflow)
hypothesized <- tribble(
  ~ purple_link, ~ prop,
  "Hello, old friend", 1 / 2,
  "Amused"           , 1 / 6,
  "Indifferent"      , 1 / 6,
  "Annoyed"          , 1 / 6
) %>%
  mutate(n = prop * n_total)
```

```
# A tibble: 4 x 3
  purple_link        prop     n
  <chr>             <dbl> <dbl>
1 Hello, old friend 0.5   1228.
2 Amused            0.167  409.
3 Indifferent       0.167  409.
4 Annoyed           0.167  409.
```

Visualizing counts

```
ggplot(purple_link_counts, aes(purple_link, n)) +
  geom_col() +
  geom_point(data = hypothesized, color = "purple")
```



chi-square goodness of fit test using chisq_test()

```
hypothesized_props <- c(
  "Hello, old friend" = 1 / 2,
  Amused              = 1 / 6,
  Indifferent         = 1 / 6,
  Annoyed             = 1 / 6
)
```

```
library(infer)
stack_overflow %>%
  chisq_test(
    response = purple_link,
    p = hypothesized_props
  )
```

```
# A tibble: 1 x 3
  statistic chisq_df       p_value
      <dbl>    <dbl>         <dbl>
1      44.0        3 0.00000000154
```

Let's practice!