Time for t

HYPOTHESIS TESTING IN R

Richie Cotton
Data Evangelist at DataCamp

Two-sample problems

- Another problem is to compare sample statistics across groups of a variable.
- `converted_comp` is a numerical variable.
- `age_first_code_cut` is a categorical variable with levels (`"child"` and `"adult"`).
- Do users who first programmed as a child tend to be compensated higher than those that started as adults?

Hypotheses

H_0: The mean compensation (in USD) is **the same** for those that coded first as a child and those that coded first as an adult.

H_0: $\mu_{child} = \mu_{adult}$

H_0: $\mu_{child} - \mu_{adult} = 0$

H_A: The mean compensation (in USD) is **greater** for those that coded first as a child compared to those that coded first as an adult.

H_A: $\mu_{child} > \mu_{adult}$

H_A: $\mu_{child} - \mu_{adult} > 0$

Calculating groupwise summary statistics

```
stack_overflow %>%
  group_by(age_first_code_cut) %>%
  summarize(mean_compensation = mean(converted_comp))
```

```
# A tibble: 2 x 2
  age_first_code_cut mean_compensation
  <chr>                          <dbl>
1 adult                        111544.
2 child                        138275.
```

Test statistics

- Sample mean estimates the population mean.
- $\bar{x}$ denotes a sample mean.
- $\bar{x}_{child}$ is the original sample mean compensation for coding first as a child.
- $\bar{x}_{adult}$ is the original sample mean compensation for coding first as an adult.
- $\bar{x}_{child} - \bar{x}_{adult}$ is a *test statistic*.
- z-scores are one type of (standardized) test statistic.

Standardizing the test statistic

$$z = \frac{\text{sample stat} - \text{population parameter}}{\text{standard error}}$$

$$t = \frac{\text{difference in sample stats} - \text{difference in population parameters}}{\text{standard error}}$$

$$t = \frac{(\bar{x}_{\text{child}} - \bar{x}_{\text{adult}}) - (\mu_{\text{child}} - \mu_{\text{adult}})}{SE(\bar{x}_{\text{child}} - \bar{x}_{\text{adult}})}$$

Standard error

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

s is the standard deviation of the variable.

n is the sample size (number of observations/rows in sample).

Assuming the null hypothesis is true

$$t = \frac{(\bar{x}_{\text{child}} - \bar{x}_{\text{adult}}) - (\mu_{\text{child}} - \mu_{\text{adult}})}{SE(\bar{x}_{\text{child}} - \bar{x}_{\text{adult}})}$$

H_0: $\mu_{\text{child}} - \mu_{\text{adult}} = 0$

$$t = \frac{(\bar{x}_{\text{child}} - \bar{x}_{\text{adult}})}{SE(\bar{x}_{\text{child}} - \bar{x}_{\text{adult}})}$$

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

```
stack_overflow %>%
  group_by(age_first_code_cut) %>%
  summarize(
    xbar = mean(converted_comp),
    s = sd(converted_comp),
    n = n()
  )
```

```
# A tibble: 2 x 4
  age_first_code_cut    xbar       s     n
  <chr>                <dbl>   <dbl> <int>
1 adult              111544. 270381.  1579
2 child              138275. 278130.  1001
```

Calculating the test statistic

```
# A tibble: 2 x 4
  age_first_code_cut    xbar       s     n
  <chr>                <dbl>   <dbl> <int>
1 adult              111544. 270381.  1579
2 child              138275. 278130.  1001
```

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

```
numerator <- xbar_child - xbar_adult
denominator <- sqrt(
  s_child ^ 2 / n_child + s_adult ^ 2 / n_adult
)
t_stat <- numerator / denominator
```

```
2.4046
```

Let's practice!

HYPOTHESIS TESTING IN R

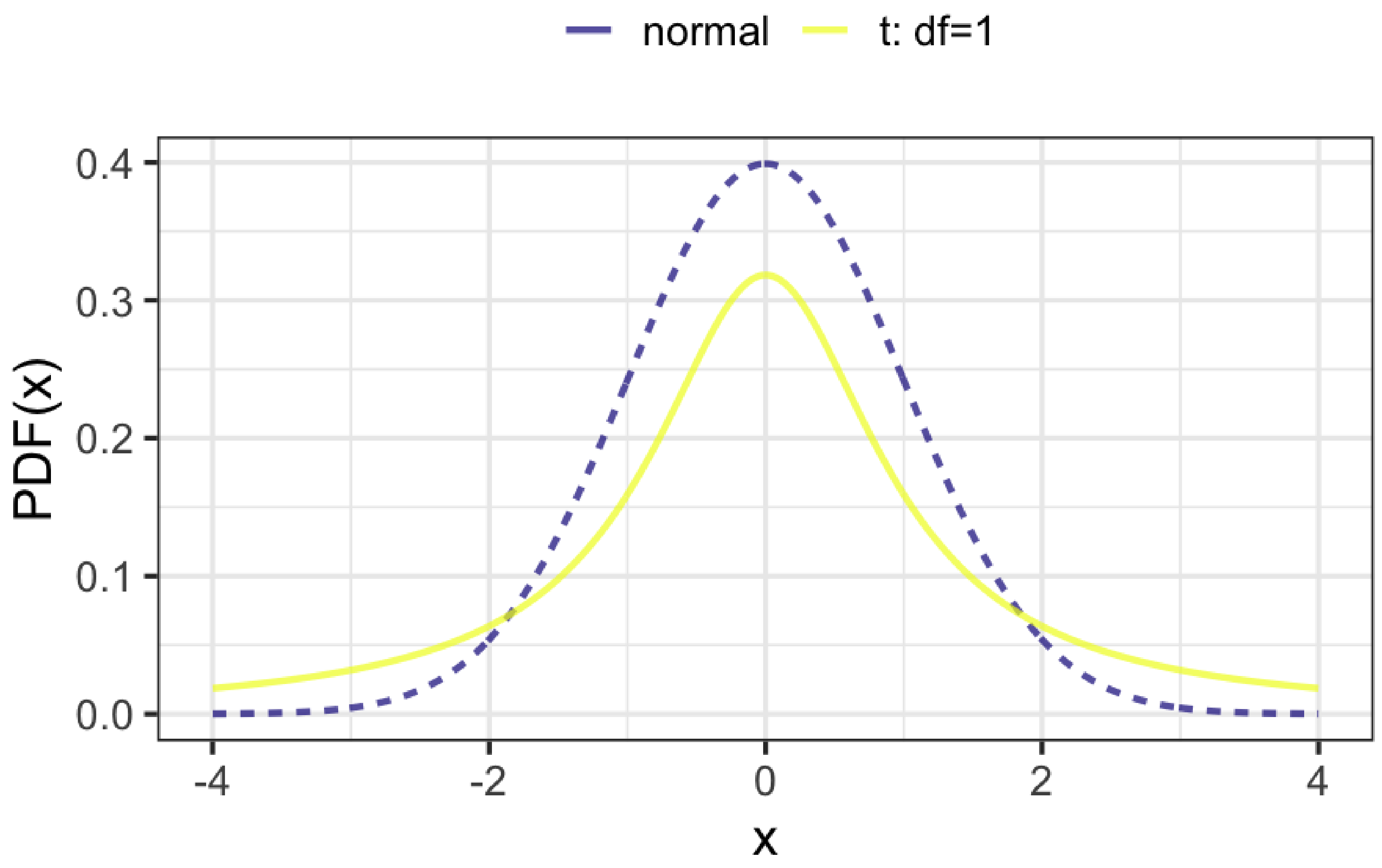
Time for t

HYPOTHESIS TESTING IN R

Richie Cotton
Data Evangelist at DataCamp

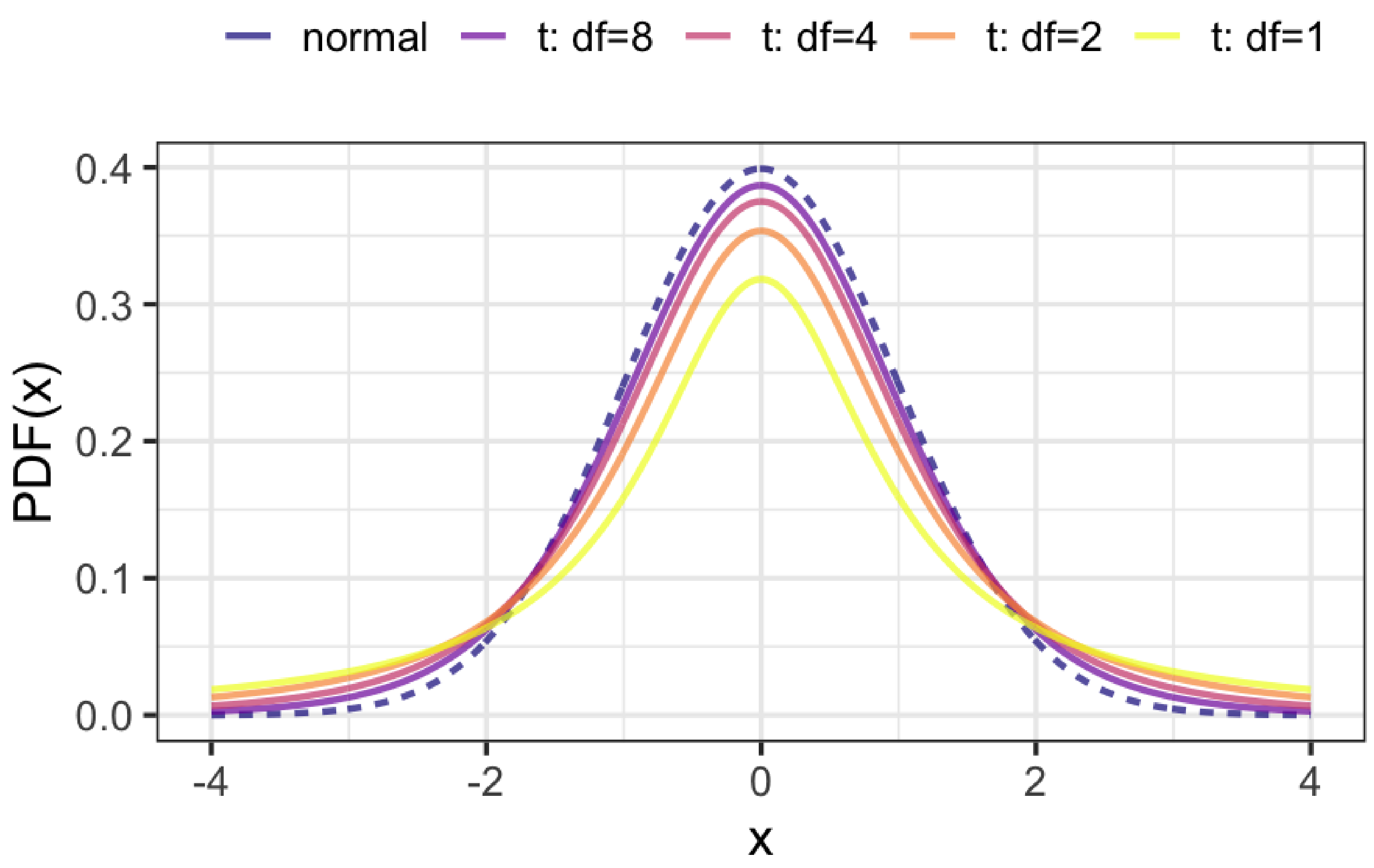
t-distributions

- The test statistic, t, follows a t-distribution.
- t-distributions have a parameter named *degrees of freedom*, or *df*.
- t-distributions look like normal distributions, with fatter tails.



Degrees of freedom

- As you increase the degrees of freedom, the t-distribution gets closer to the normal distribution.
- A normal distribution is a t-distribution with infinite degrees of freedom.
- Degrees of freedom are the maximum number of logically independent values in the data sample.



Calculating degrees of freedom

- Suppose your dataset has 5 independent observations.
- Four of the values are 2, 6, 8, and 5.
- You also know the sample mean is 5.
- The last value is no longer independent; it must be 4.
- There are 4 degrees of freedom.
- $df = n_{child} + n_{adult} - 2$

Hypotheses

H_0: The mean compensation (in USD) is **the same** for those that coded first as a child and those that coded first as an adult.

H_A: The mean compensation (in USD) is **greater** for those that coded first as a child compared to those that coded first as an adult.

Use a **right-tailed test**.

Significance level

$\alpha = 0.1$

If $p \leq \alpha$ then reject H_0.

Calculating p-values: one proportion vs. a value

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

Calculating p-values: two means from different groups

```
numerator <- xbar_child - xbar_adult
denominator <- sqrt(s_child ^ 2 / n_child + s_adult ^ 2 / n_adult)
t_stat <- numerator / denominator
```

```
2.4046
```

```
degrees_of_freedom <- n_child + n_adult - 2
```

```
2578
```

- Test statistic standard error used an approximation (not bootstrapping).
- Use t-distribution CDF not normal CDF.

```
p_value <- pt(t_stat, df = degrees_of_freedom, lower.tail = FALSE)
```

```
0.008130
```

Let's practice!

HYPOTHESIS TESTING IN R

Pairing is caring

HYPOTHESIS TESTING IN R

Richie Cotton
Data Evangelist at DataCamp

US Republican presidents dataset

state	county	repub_percent_08	repub_percent_12
Alabama	Bullock	25.69	23.51
Alabama	Chilton	78.49	79.78
Alabama	Clay	73.09	72.31
Alabama	Cullman	81.85	84.16
Alabama	Escambia	63.89	62.46
Alabama	Fayette	73.93	76.19
Alabama	Franklin	68.83	69.68
...	...	...	...

500 rows; each row represents county-level votes in a presidential election.

[1] https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/VOQCHQ

Hypotheses

Question: Was the percentage of votes given to the Republican candidate lower in 2008 compared to 2012?

H_0: $\mu_{2008} - \mu_{2012} = 0$

H_A: $\mu_{2008} - \mu_{2012} < 0$

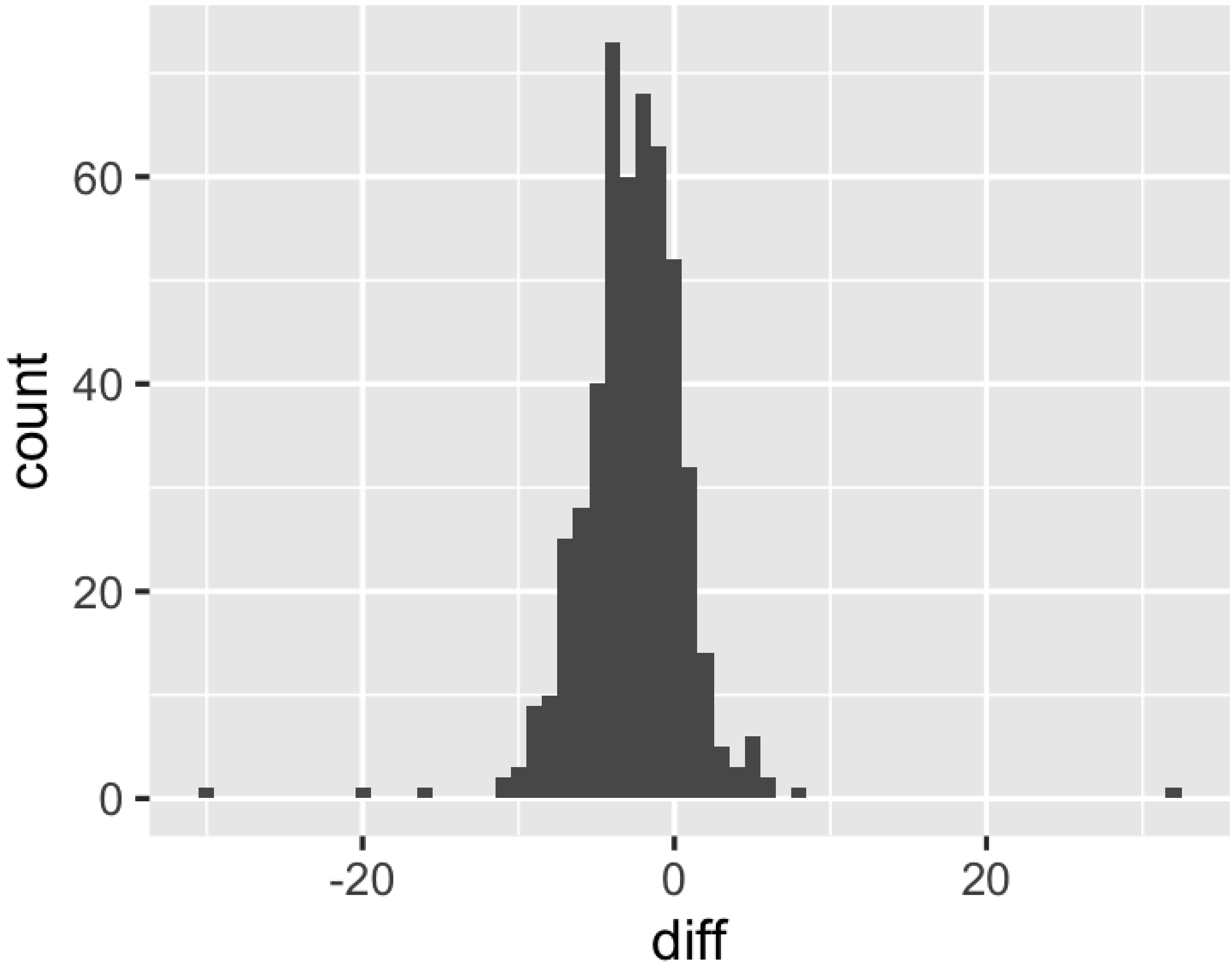
Set $\alpha = 0.05$ significance level.

The data is paired, since each voter percentage refers to the same county.

From two samples to one

```
sample_data <- repub_votes_potus_08_12 %>%
  mutate(diff = repub_percent_08 - repub_percent_12)
```

```
ggplot(sample_data, aes(x = diff)) +
  geom_histogram(binwidth = 1)
```



Calculate sample statistics of the difference

```
sample_data %>%
  summarize(xbar_diff = mean(diff))
```

```
  xbar_diff
1 -2.643027
```

Revised hypotheses

Old hypotheses

H_0: $\mu_{2008} - \mu_{2012} = 0$

H_A: $\mu_{2008} - \mu_{2012} < 0$

New hypotheses

H_0: $\mu_{\text{diff}} = 0$

H_A: $\mu_{\text{diff}} < 0$

$$t = \frac{\bar{x}_{\text{diff}} - \mu_{\text{diff}}}{\sqrt{\frac{s_{diff}^2}{n_{\text{diff}}}}}$$

$$df = n_{diff} - 1$$

Calculating the p-value

```
n_diff <- nrow(sample_data)
```

```
s_diff <- sample_data %>%
  summarize(sd_diff = sd(diff)) %>%
  pull(sd_diff)
```

```
t_stat <- (xbar_diff - 0) / sqrt(s_diff ^ 2 / n_diff)
```

```
-16.06374
```

```
degrees_of_freedom <- n_diff - 1
```

```
499
```

$$t = \frac{\bar{x}_{\text{diff}} - \mu_{\text{diff}}}{\sqrt{\frac{s_{\text{diff}}^2}{n_{\text{diff}}}}}$$

$$df = n_{\text{diff}} - 1$$

```
p_value <- pt(t_stat, df = degrees_of_freedom)
```

```
2.084965e-47
```

Testing differences between two means using t.test()

```
t.test(
  # Vector of differences
  sample_data$diff,
  # Choose between "two.sided", "less", "greater"
  alternative = "less",
  # Null hypothesis population parameter
  mu = 0
)
```

```
	One Sample t-test

data:  sample_data$diff
t = -16.064, df = 499, p-value < 2.2e-16
alternative hypothesis: true mean is less than 0
95 percent confidence interval:
     -Inf -2.37189
sample estimates:
mean of x
-2.643027
```

t.test() with paired = TRUE

```r
t.test(
  sample_data$repub_percent_08,
  sample_data$repub_percent_12,
  alternative = "less",
  mu = 0,
  paired = TRUE
)
```

```
	Paired t-test

data:  sample_data$repub_percent_08 and
       sample_data$repub_percent_12
t = -16.064, df = 499, p-value < 2.2e-16
alternative hypothesis: true difference in means
                        is less than 0
95 percent confidence interval:
     -Inf -2.37189
sample estimates:
mean of the differences
              -2.643027
```

Unpaired t.test()

```
t.test(
  x = sample_data$repub_percent_08,
  y = sample_data$repub_percent_12,
  alternative = "less",
  mu = 0
)
```

Unpaired t-test has more chance of false negative error (less statistical power).

```
	Welch Two Sample t-test

data:  sample_data$repub_percent_08 and
       sample_data$repub_percent_12
t = -2.8788, df = 992.76, p-value = 0.002039
alternative hypothesis: true difference in means
                        is less than 0
95 percent confidence interval:
      -Inf -1.131469
sample estimates:
mean of x mean of y
 56.52034  59.16337
```

Let's practice!

HYPOTHESIS TESTING IN R

P-hacked to pieces

HYPOTHESIS TESTING IN R

Richie Cotton
Data Evangelist at DataCamp

Job satisfaction: 5 categories

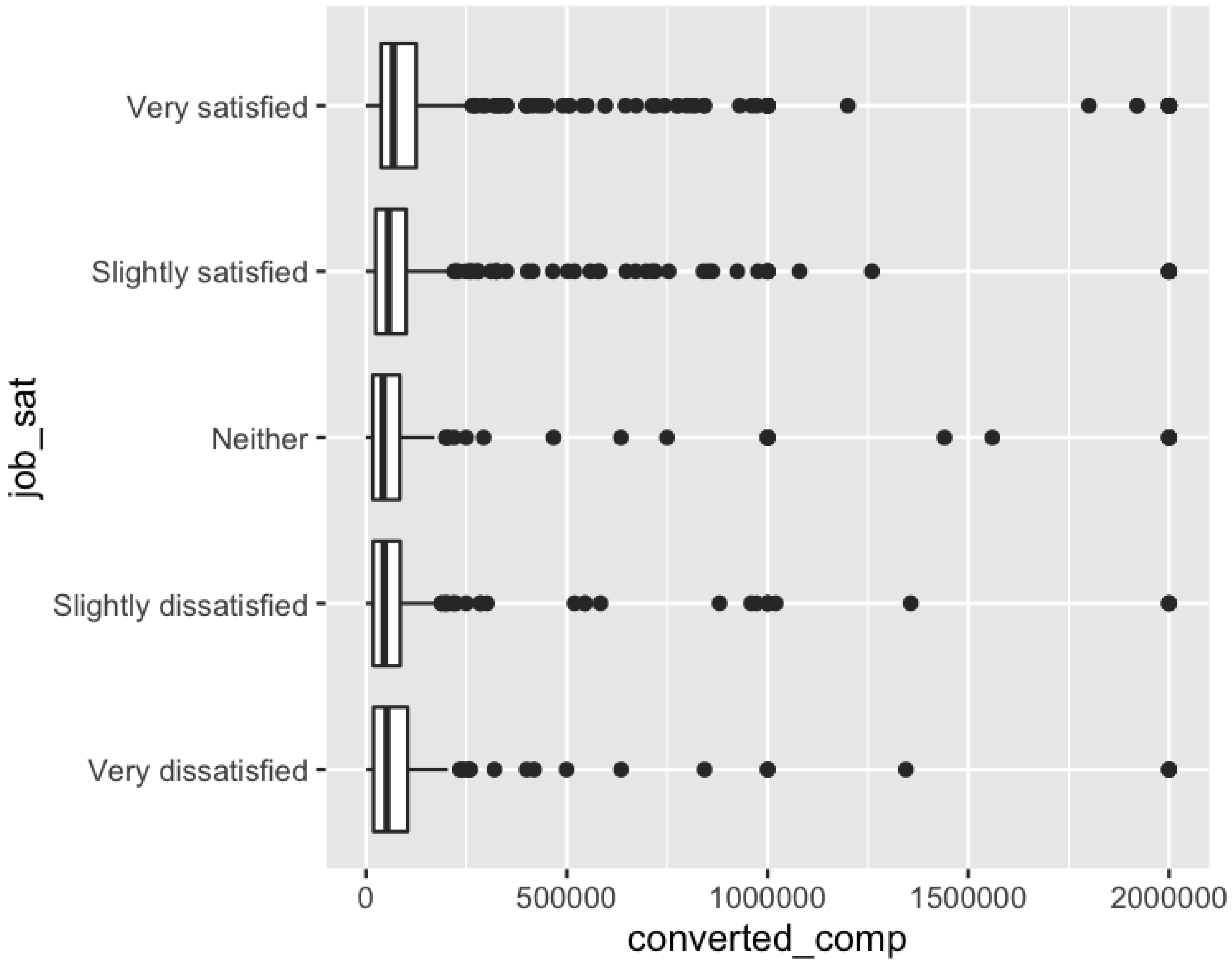
```
stack_overflow %>%
  count(job_sat)
```

```
# A tibble: 5 x 2
  job_sat                   n
  <fct>                 <int>
1 Very dissatisfied       187
2 Slightly dissatisfied   385
3 Neither                 245
4 Slightly satisfied      777
5 Very satisfied          981
```

Visualizing multiple distributions

Question: Is mean annual compensation different for different levels of job satisfaction?

```
stack_overflow %>%
  ggplot(aes(x = job_sat, y = converted_comp)) +
  geom_boxplot() +
  coord_flip()
```



Analysis of variance (ANOVA)

```
mdl_comp_vs_job_sat <- lm(converted_comp ~ job_sat, data = stack_overflow)
```

```
anova(mdl_comp_vs_job_sat)
```

```
Analysis of Variance Table


Response: converted_comp
            Df   Sum Sq  Mean Sq F value Pr(>F)
job_sat      4 1.09e+12 2.73e+11    3.65 0.0057 **
Residuals 2570 1.92e+14 7.47e+10


Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

[1] Linear regressions with lm() are taught in "Introduction to Regression in R"

Pairwise tests

- $\mu_{\text{very dissatisfied}} \neq \mu_{\text{slightly dissatisfied}}$
- $\mu_{\text{very dissatisfied}} \neq \mu_{\text{neither}}$
- $\mu_{\text{very dissatisfied}} \neq \mu_{\text{slightly satisfied}}$
- $\mu_{\text{very dissatisfied}} \neq \mu_{\text{very satisfied}}$
- $\mu_{\text{slightly dissatisfied}} \neq \mu_{\text{neither}}$
- $\mu_{\text{slightly dissatisfied}} \neq \mu_{\text{slightly satisfied}}$
- $\mu_{\text{slightly dissatisfied}} \neq \mu_{\text{very satisfied}}$
- $\mu_{\text{neither}} \neq \mu_{\text{slightly satisfied}}$
- $\mu_{\text{neither}} \neq \mu_{\text{very satisfied}}$
- $\mu_{\text{slightly satisfied}} \neq \mu_{\text{very satisfied}}$

Set significance level to $\alpha = 0.2$.

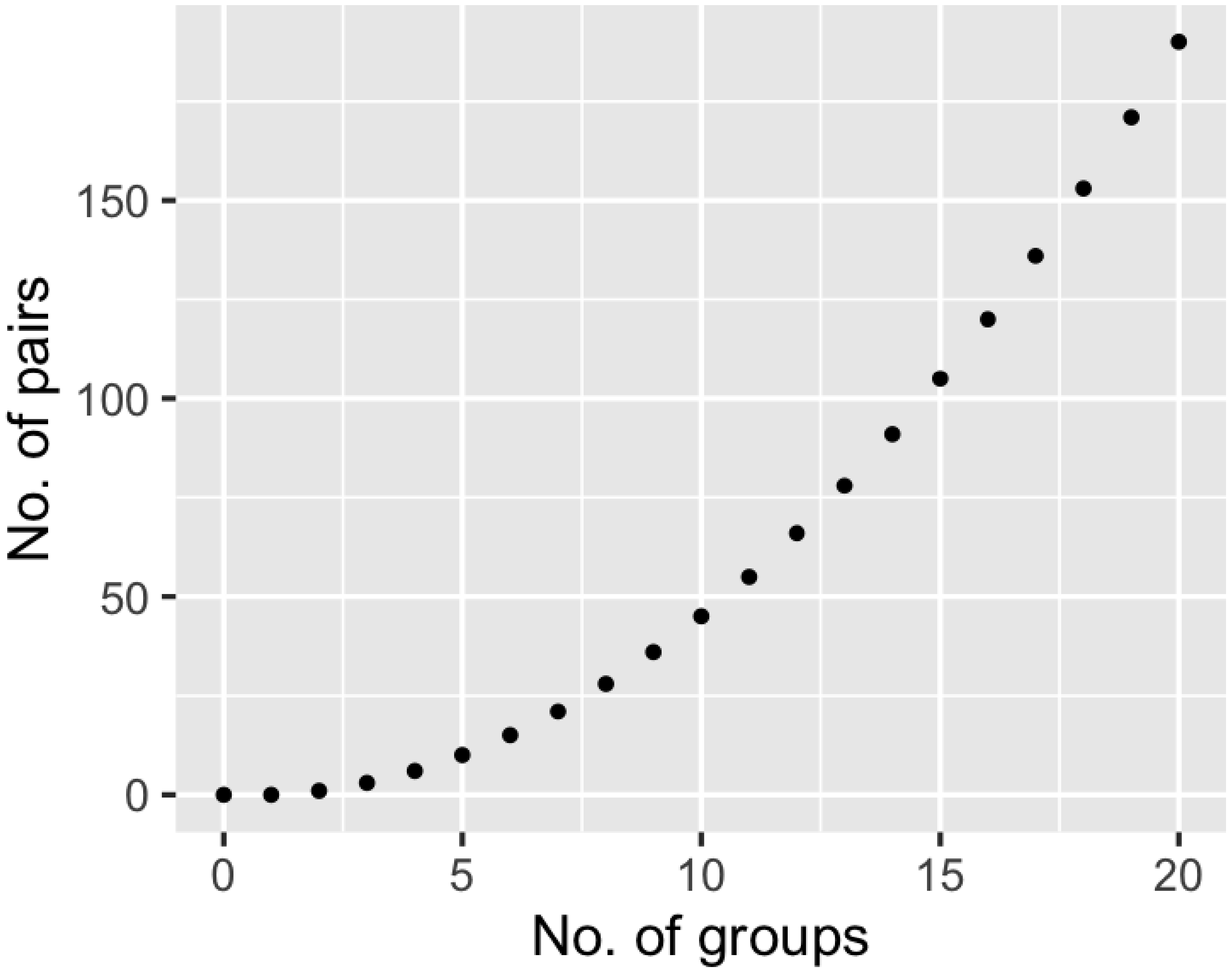
pairwise.t.test()

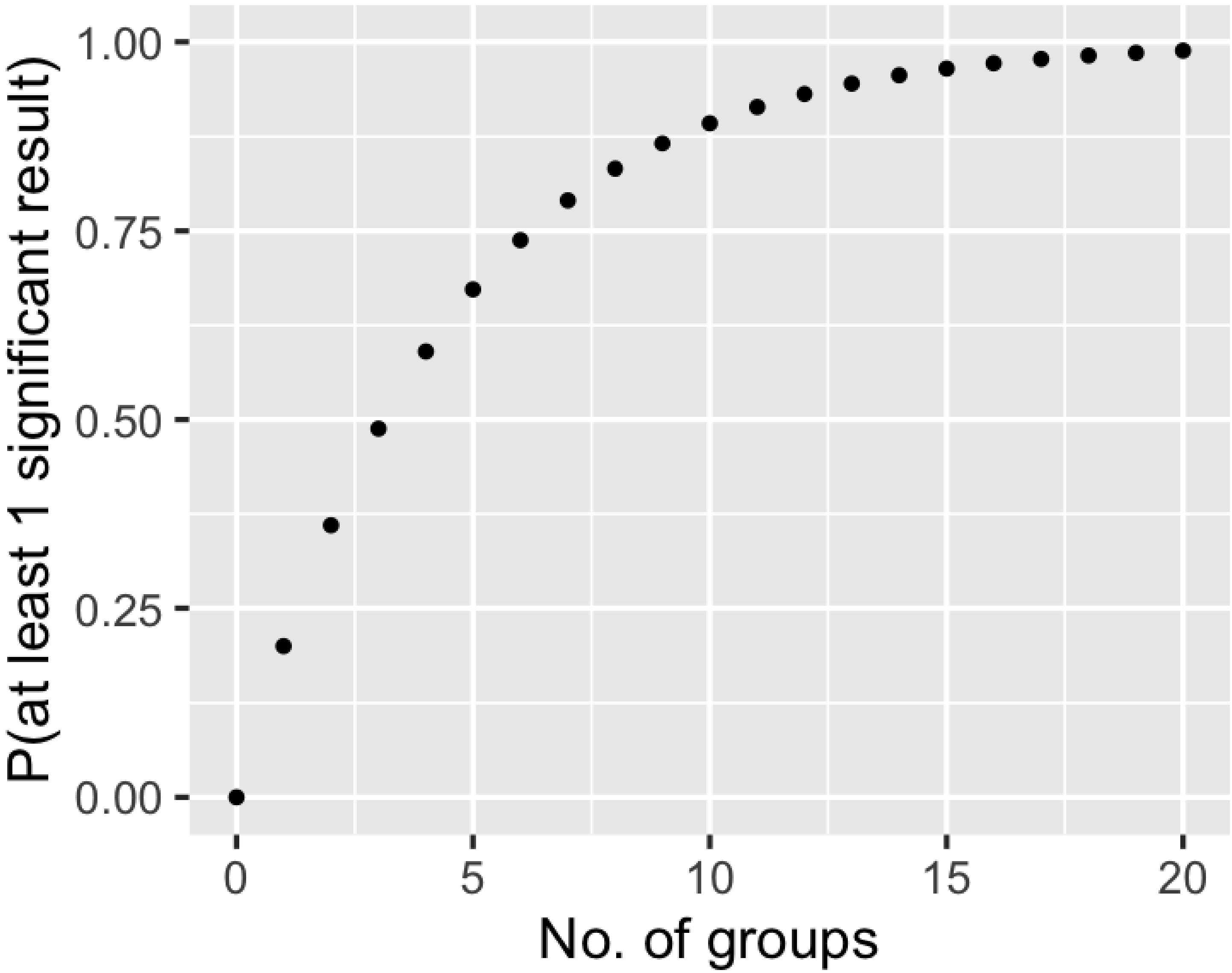
```
pairwise.t.test(stack_overflow$converted_comp, stack_overflow$job_sat, p.adjust.method = "none")
```

```
	Pairwise comparisons using t tests with pooled SD 

data:  stack_overflow$converted_comp and stack_overflow$job_sat 

                      Very dissatisfied Slightly dissatisfied Neither Slightly satisfied
Slightly dissatisfied 0.26860           -                     -       -                 
Neither               0.79578           0.36858               -       -                 
Slightly satisfied    0.29570           0.82931               0.41248 -                 
Very satisfied        0.34482           0.00384               0.15939 0.00084           

P value adjustment method: none
```

Significant differences: "Very satisfied" vs. "Slightly dissatisfied"; "Very satisfied" vs. "Neither"; "Very satisfied" vs. "Slightly satisfied"

As the no. of groups increases...





Bonferroni correction

```
pairwise.t.test(stack_overflow$converted_comp, stack_overflow$job_sat, p.adjust.method = "bonferroni")
```

```
    Pairwise comparisons using t tests with pooled SD

data:  stack_overflow$converted_comp and stack_overflow$job_sat

                      Very dissatisfied Slightly dissatisfied Neither Slightly satisfied
Slightly dissatisfied 1.0000            -                     -       -
Neither               1.0000            1.0000                -       -
Slightly satisfied    1.0000            1.0000                1.0000  -
Very satisfied        1.0000            0.0384                1.0000  0.0084

P value adjustment method: bonferroni
```

Significant differences: "Very satisfied" vs. "Slightly dissatisfied"; "Very satisfied" vs. "Slightly satisfied"

More methods

```
p.adjust.methods
```

```
"holm"  "hochberg"  "hommel"  "bonferroni"  "BH"  "BY"  "fdr"  "none"
```

Bonferroni and Holm adjustments

```
p_values
```

```
0.268603 0.795778 0.295702 0.344819 0.368580 0.829315 0.003840 0.412482 0.159389 0.000838
```

Bonferroni

```
pmin(1, 10 * p_values)
```

```
1.00000 1.00000 1.00000 1.00000 1.00000 1.00000 0.03840 1.00000 1.00000 0.00838
```

Holm (roughly)

```
pmin(1, 10:1 * sort(p_values))
```

```
0.00838 0.03456 1.00000 1.00000 1.00000 1.00000 1.00000 1.00000 1.00000 0.82931
```

Let's practice!

HYPOTHESIS TESTING IN R