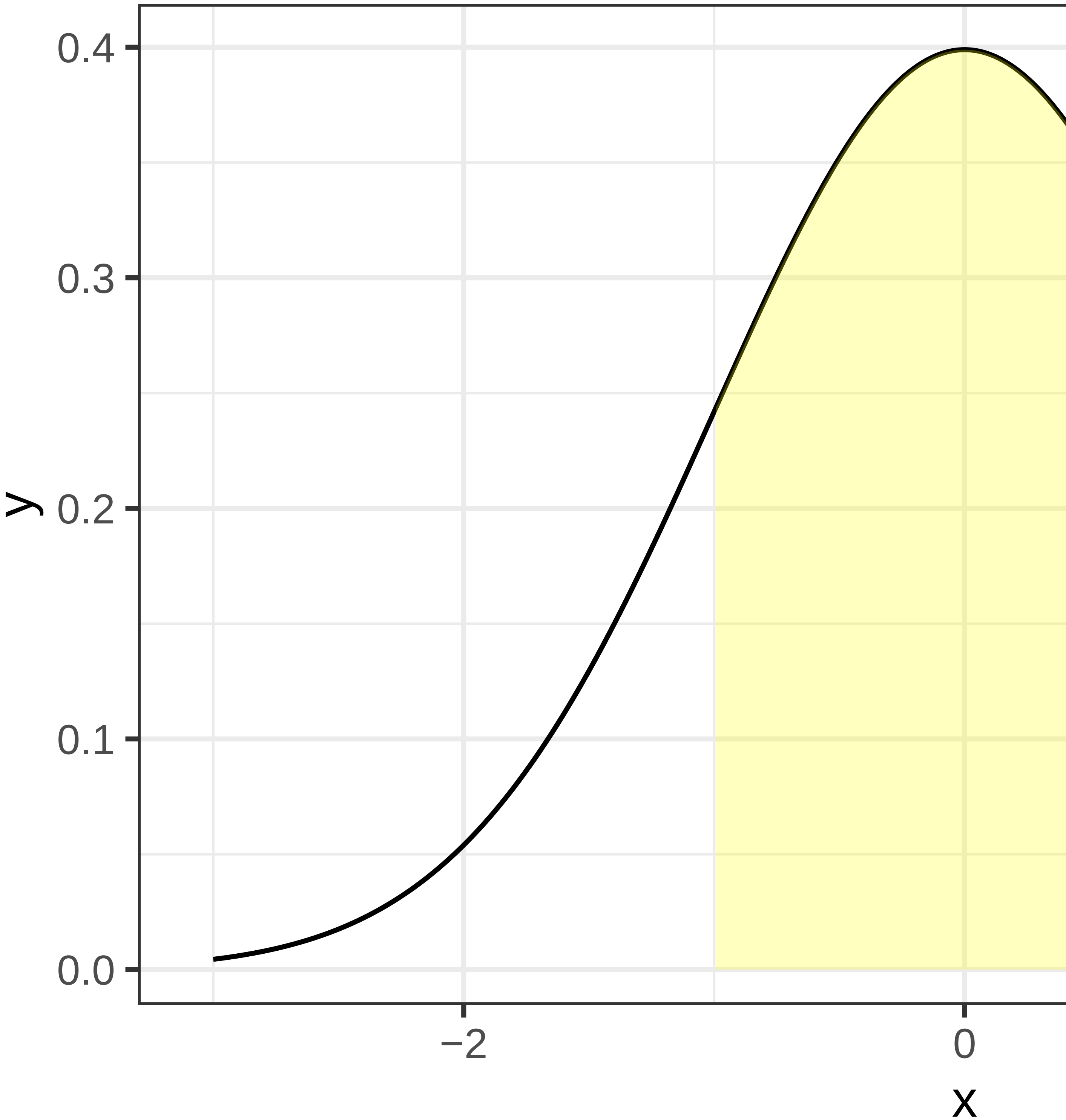
Lecture 4

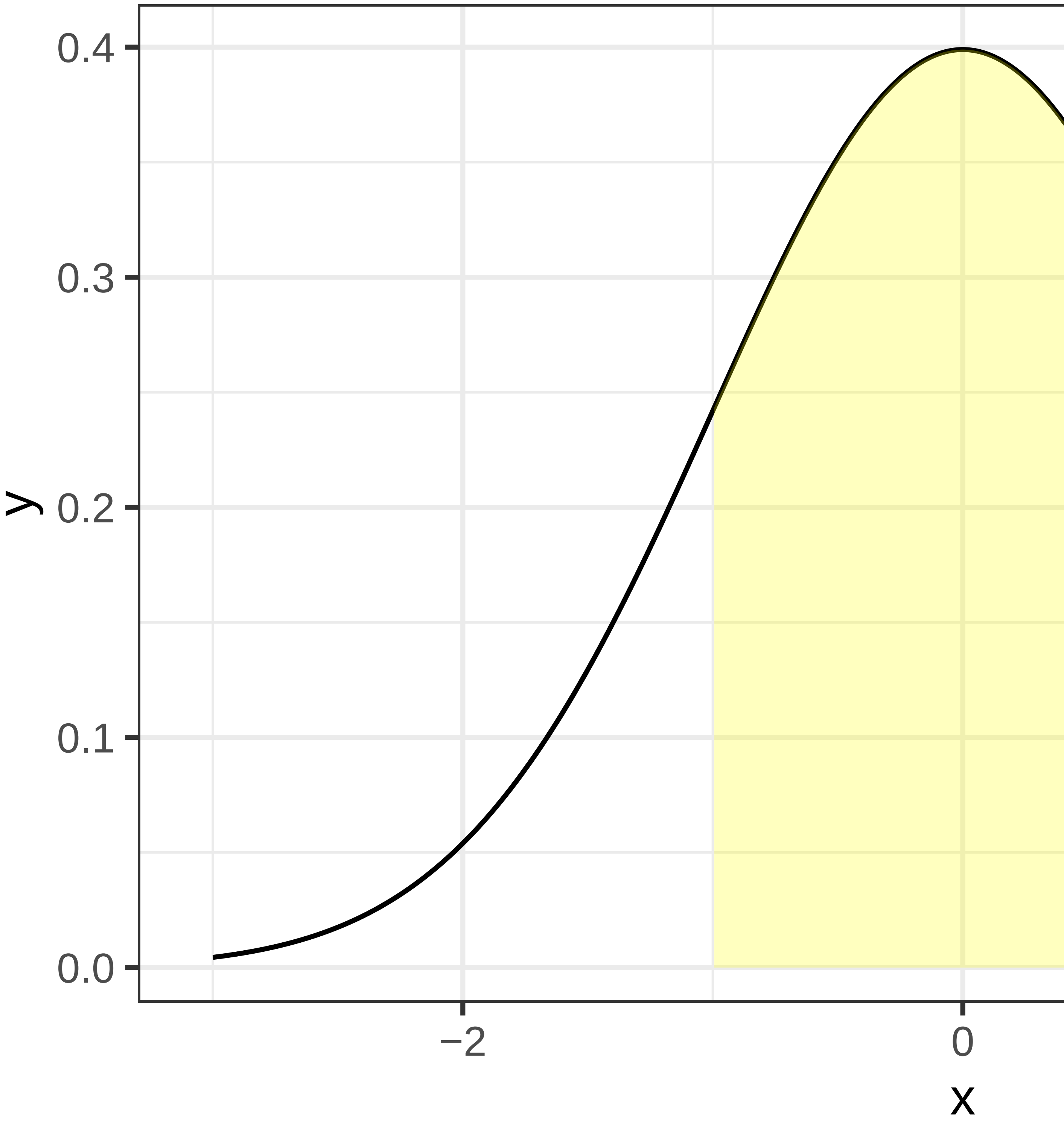
Hypothesis testing: motivating question

- Run $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$
- Suppose $\beta_1 = 0$ (the true value), will we obtain $\hat{\beta}_1 = 0$?
- No. Even without bias we will still have some random variation due to sampling
- We know that $\hat{\beta}_1 \sim N(\beta_1, \frac{\hat{\sigma}_\varepsilon}{n\sigma_x^2})$
  - Assuming Gauss-Markov assumptions

Probabilities under continuous random variables

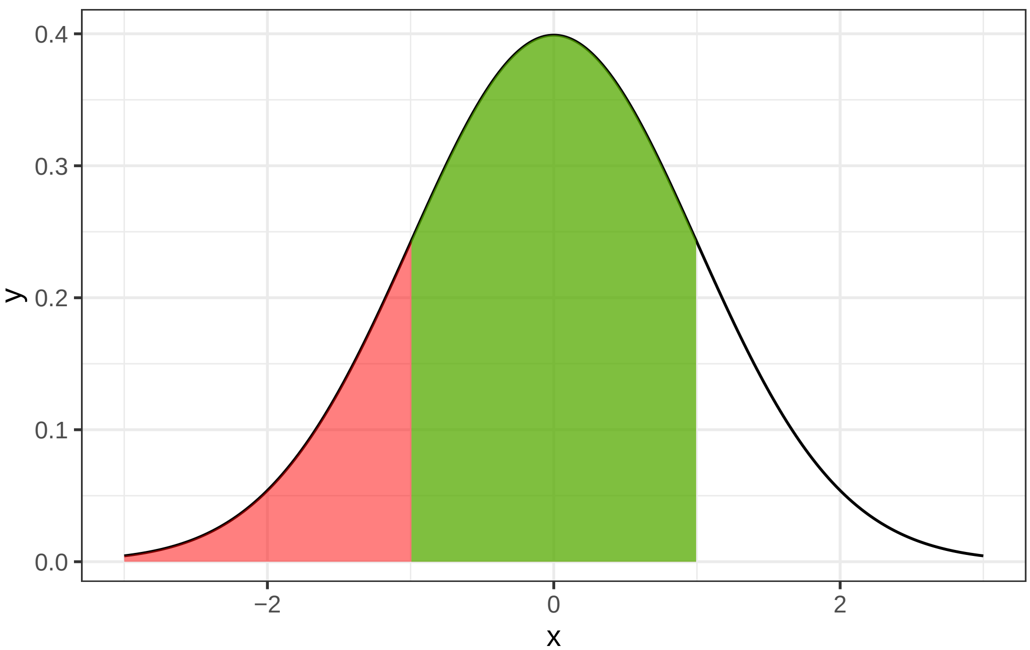


- This is a probability density function
- What is the probability that X=0?
    - It's exactly 0


0.4
0.3
0.2
0.1
0.0
y
−2
0
x

- We want to find the area of the shaded region (between -1 and 1)
- The function pnorm ($= \Phi(z)$) gives $P(X \leq z)$
- How do we calculate $P(-1 \leq x \leq 1)$?

Continuous RV



- We grab the full area to the left of 1 (pnrom(1) = green + red area), then subtract out the area to the left of -1 (pnorm(-1))

Question: Probabilities

- $\hat{\beta}_1 \sim N(0,1)$. What is the probability that $\hat{\beta}_1$ is between -2 and 1, $P(-2 < \hat{\beta}_1 < 1)$?

```
      x area_to_left
1: -3.0        0.001
2: -2.5        0.006
3: -2.0        0.023
4: -1.5        0.067
5: -1.0        0.159
6: -0.5        0.309
7:  0.0        0.500
8:  0.5        0.691
```

```
 9:  1.0         0.841
10:  1.5         0.933
11:  2.0         0.977
12:  2.5         0.994
13:  3.0         0.999
```

Answer: Probabilities

```
pnorm(1)-pnorm(-2)
```

```
[1] 0.8185946
```

Question: calculating z scores

- Given a standard normal ($Z \sim N(0,1)$) we can calculate $P(a < X < b)$ with $\Phi(b) - \Phi(a)$ where $\Phi(x)$ is the pnorm function we used earlier
- If $X \sim N(\mu, \sigma)$ how we calculate $P(a < X < b)$?
- Standardize: subtract the mean and divide by the standard deviation to obtain the z score. This is still normal.
- $P(a < x < b) = P(\frac{a-\mu}{\sigma} < \frac{X-\mu}{\sigma} < \frac{b-\mu}{\sigma}) = P(z_a < Z < z_b)$
- $= \Phi(z_a) - \Phi(z_b)$

Question: z scores normal

- what is the probability that $\hat{\beta}_1$ is between -2 and 2, but now $\hat{\beta}_1 \sim N(1, 2)$, ie it now has mean 1 and standard error of 2?

```
      x area_to_left
 1: -3.0         0.001
 2: -2.5         0.006
 3: -2.0         0.023
 4: -1.5         0.067
 5: -1.0         0.159
 6: -0.5         0.309
 7:  0.0         0.500
 8:  0.5         0.691
 9:  1.0         0.841
```

```
10:  1.5         0.933
11:  2.0         0.977
12:  2.5         0.994
13:  3.0         0.999
```

Answer: z scores normal

- $(\hat{\beta}_1 - 1)/2 \equiv z \sim N(0,1)$
- Now translate the probability: $-2 < \beta_2 < 2 \implies -1.5 < z < .5$

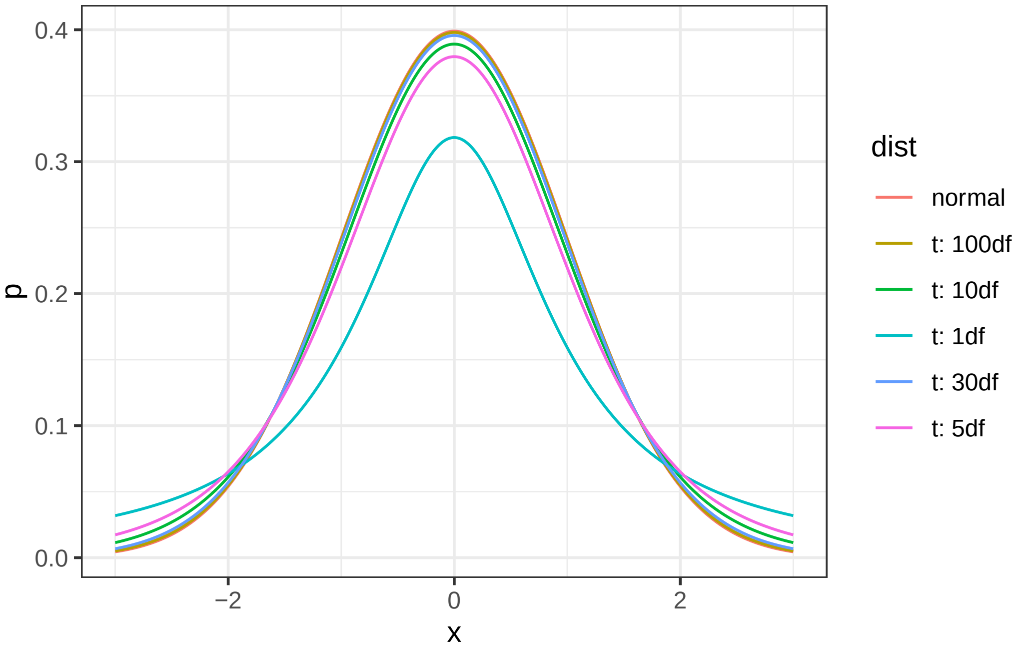
```
        x area_to_left
 1: -3.0         0.001
 2: -2.5         0.006
 3: -2.0         0.023
 4: -1.5         0.067
 5: -1.0         0.159
 6: -0.5         0.309
 7:  0.0         0.500
 8:  0.5         0.691
 9:  1.0         0.841
10:  1.5         0.933
11:  2.0         0.977
12:  2.5         0.994
13:  3.0         0.999
```

```
[1] 0.6246553
```

t distribution vs Z

- Asymptotically (as $n \to \infty$) $\hat{\beta}_1$ is normally distributed, but for small n it actually follows a t distribution
- The t distribution has a mean and standard deviation, but also degrees of freedom ($df = n - 1$)
- Nothing actually changes, except when we lookup a table or calculate in R we use slightly different function to calculate
- In R: pt(z,df) instead of pnorm(z)
- It is rare to have small samples in econometrics - they basically give the same result
- R regression output gives you t values by default

T distribution: graph



Flipping the question

- We know that if $\hat{\beta}_1 = 0$ we may still get nonzero results
- If we obtain $\hat{\beta}_1 = 1$, what is the probability that $\beta_1 = 0$? i.e. what is the probability there is actually no effect but we obtain a nonzero result?
- This is impossible to answer, but we can get a suggestive answer from the prior exercise
- We assume the null hypothesis ($\beta_1 = 0$), and calculate the probability that we obtain a value of $\hat{\beta}_1 \geq 1$, ie $\hat{\beta}_1 \sim N(0,1), p = P(\hat{\beta}_1 > 1) = 1 - \Phi(1) \approx .32$,

Adding Hypothesis Testing Jargon

- We are testing the hypothesis that x causes y, ie that $\beta_1 \neq 0$ (we are explicitly assuming exogeneity at the moment). We label these hypotheses.
  - H_0 is always the null hypothesis: that no effect exists. Here $H_0 : \beta_1 = 0$
  - H_1 is the alternative hypothesis that there is an effect: $H_1 : \beta_1 \neq 0$. We could have also explicitly tested $\beta_1 > 0$ or $\beta_1 < 0$

Adding Hypothesis Testing Jargon

- We don't have enough information to calculate probabilities. Instead we ask: if H_0 is true, what is the probability that we observe a result as extreme as $\hat{\beta}_1$? Suppose we obtained an estimate of $\hat{\beta}_1 = 1$
  - $P(|\hat{\beta}_1| > 1 \mid \beta_1 = 0)$
  - This is called the p value, here $p = 1 - .68 = .32$

Adding Hypothesis Testing Jargon

- We set up a criterion for rejecting H_0 which we call the significance level α (normally .05 or .01). If the probability of obtaining such a result under the null is small, we reject the null hypothesis, otherwise we conservatively fail to reject H_0
  - Here $\alpha = .05 > p = .32$ so we fail to reject H_0
  - There could still be an effect (on average it's 1), it's just too small to confidently conclude it wasn't due to random chance

iClicker

You run a regression and obtain $\hat{\beta}_1 = 0.01$ with a p value of 0.001. Which of the following conclusions is the most correct?

- A There is a 0.1 percent chance that there is no effect of x on y
- B There is a small effect of x on y
- C A null effect with our given model is inconsistent with the data
- D There is an effect of x on y, but we don't know how large

iClicker

You run a regression and obtain $\hat{\beta}_1 = 1$ with a p value of 0.5. Which of the following conclusions is the most correct?

- A There is no effect of x on y
- B We cannot rule out a null effect of x on y
- C There is only a 50 percent chance that there is an effect of x on y
- D If there is an effect of x on y, it must be a small effect

Adding Hypothesis Testing Jargon

- Under this setup, we will reject $\alpha = 5\%$ of cases where $\beta_1 = 0$ by chance (the type I error rate, or false positive rate)
- We can also fail to reject H_0 (and conclude there is no effect) even when $\beta_1 \neq 0$, called a type II error or false negative
  - This occurs with probability β, which requires additional assumptions to calculate

Hypothesis testing: Type 1 and Type 2 errors

- We can either reject or fail to reject H_0, additional H_0 can either be true or false (note that we can never observe this). This leads to 4 possibilities

Hypothesis testing: Type 1 and Type 2 errors

- H_0 is true and you fail to reject H_0 : **correct decision**
- H_0 is true and you reject H_0: **False positive (type 1 error)** This happens with probability α
- H_0 is false and you fail to reject H_0: **False negative (type 2 error)**. This happens with probability β, which requires additional assumptions to calculate
- H_0 is false and you reject H_0 **Correct decision**

Hypothesis Testing: Example

- We randomly assign students to classroom with either 10 or 20 students in an experiment to determine the effect of class size on test scores. We run the regression $score_i = \beta_0 + \beta_1 size_i + \varepsilon_i$. After running an OLS regression we obtain $\hat{\beta}_1 = 0.05, se(\hat{\beta}_1) = 0.04$ We wish to know whether class size affects student test scores, and use a significance level of $\alpha = .05$

- Write H_0, H_1 using a two tailed test
- Calculate the standardized values (z scores) for the area you're testing

Hypothesis Testing: Example

- We randomly assign students to classroom with either 10 or 20 students in an experiment to determine the effect of class size on test scores. We run the regression $score_i = \beta_0 + \beta_1 size_i + \varepsilon_i$. After running an OLS regression we obtain $\hat{\beta}_1 = 0.05, se(\hat{\beta}_1) = 0.04$ We wish to know whether class size affects student test scores, and use a significance level of $\alpha = .05$

- Compute the p value
- Determine your decision

Confidence Intervals

- Suppose we reject H_0 and have $\hat{\beta}_1 = 1, se(\hat{\beta}_1) = 0.1$.
- We know that there is likely an effect (if all assumptions are met), and that the most likely value of β_1 is 1
- We would like to know the likely range of values β_1 could actually be
- Instead of testing a hypothesis, we can form an interval around $\hat{\beta}_1$ that captures 95% (or 99%, etc) of the likely range of values it can take on.

Confidence Intervals

- The calculation is essentially the same, and we call it a confidence interval: $CI_{0.95} = \hat{\beta}_1 \pm z_{.95} se(\hat{\beta}_1)$
  - $-1.96 < z_{.95} < 1.96$ gives an area of 95%, so we have $\hat{\beta}_1 = 1 \pm .196 = (.804, 1.196)$
  - How do we obtain 1.96? qnorm(.975). $.975 = 1 - .05/2$, ie we take α, divide between our two tails, and lookup the value in our table

Confidence Intervals

- We say that we are 95% confident that the true value of β_1 is between .804 and 1.196
  - We don't say that there's a 95% chance that β_1 is in this range because it isn't true. We're making a lot of assumptions when calculating this value
  - More technically, if we were to repeat the experiment many times, the 95% confidence interval we calculate during that experiment would capture the true value of β_1 95% of the time, conditional on all modeling assumptions

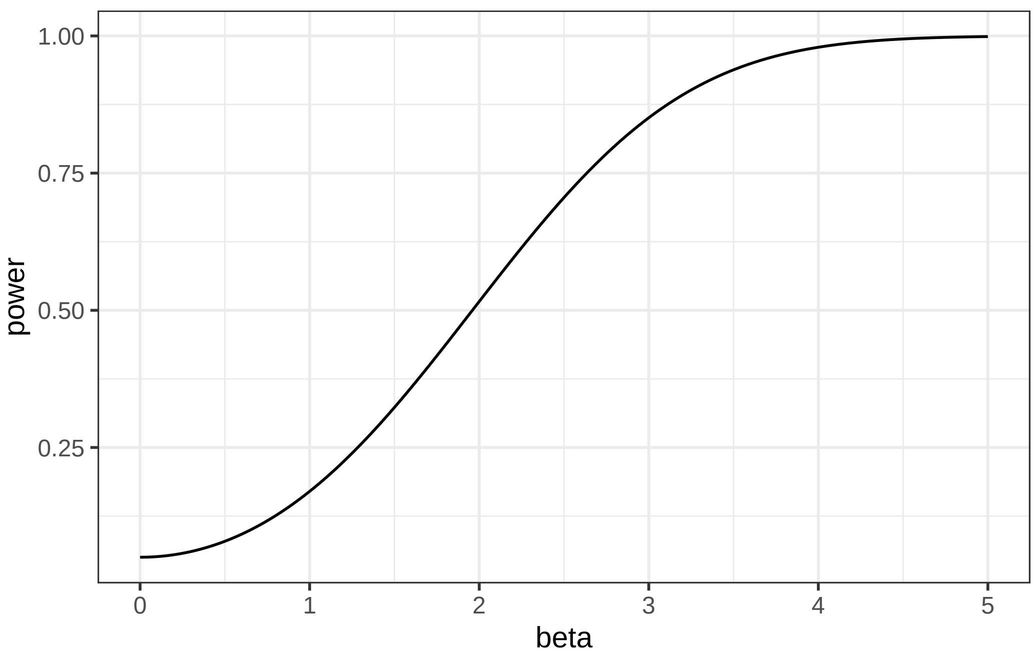
Confidence Intervals: Example

- $\hat{\beta}_1 = 3, se(\hat{\beta}_1) = 2$ calculate a 99% confidence interval of $\hat{\beta}_1$

Power Curve

- Let $\alpha = .05$, $H_0 : \hat{\beta}_1 \sim N(0, 1)$. Then we fail to reject H_0 if $-1.96 < \hat{\beta}_1 < 1.96$
- Probability of a type I error is .05. Type 2 error depends on what β_1 is
- Calculate this for each value of β_1
- Power curve graphs the "power" - $1 - \beta_1$ (so that a power of 1 corresponds to a type II error rate of 0)
- Because this is symmetric we normally only graph the right hand side (which we can reference as $|\beta_1|$)

Power Curve: Graph



Power: Implications

- Suppose my outcome variable is number of years of education achieved. A result of .0001 years may end up being statistically significant, but not practically

- Suppose I determine that anything below 1 year is not very significant. Then I can look up $\beta = 1$ in the power curve and determine my probability of detecting an effect of that size
- In the prior graph I only have a 20% chance of being able to detect such a change. I need an effect of around 3 just to have an 80% chance of detecting an event

Power: Implications

- Given power, we can now interpret results of hypothesis testing. Assuming all of our modeling assumptions hold, then:
- If we reject H_0 then we know this is likely a real effect, though we must interpret it since it could be incredibly small in practice
- If we fail to reject H_0 then it may have just been random noise, but there also could have been a sizeable effect that was missed due to lower power (e.g. $\beta = 1$)

Power: Implications

- If we have a large sample size then our interpretation is usually clear: we know whether there is an effect, and we can judge the size based on practical significance
- This is conditional on all other assumptions! This is why economic papers focus most of their effort on methodology and very little on p values

iclicker

You obtain $\hat{\beta}_1 = 15$ with a p value of 0.2. Suppose that any value above 100 is considered moderate and any value above 10 is considered small. You have 95% power to detect an effect size of $\hat{\beta}_1 = 20$. Your context is a valid randomized control trial. What can you conclude?

- A We cannot reject that there is no effect, but we also cannot rule out a moderate effect size
- B We can confidently rule out even small effects of x on y
- C We can confidently rule out moderate effects of x on y. Small effects are still possible, but we cannot reject that the effect size is zero.

Errors: Examples; categorization

- You are given the following output from R on a regression of y vs x. Do you reject $H_0 : \beta_1 = 0$ at $\alpha = .05$?

```
Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-2.6425 -0.5822 -0.1061  0.7608  2.2235

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.04081    0.10041   0.406    0.685
x            0.11418    0.09292   1.229    0.222

Residual standard error: 0.9957 on 98 degrees of freedom
Multiple R-squared:  0.01517,   Adjusted R-squared:  0.005125
F-statistic:  1.51 on 1 and 98 DF,  p-value: 0.2221
```

When Hypothesis testing goes wrong

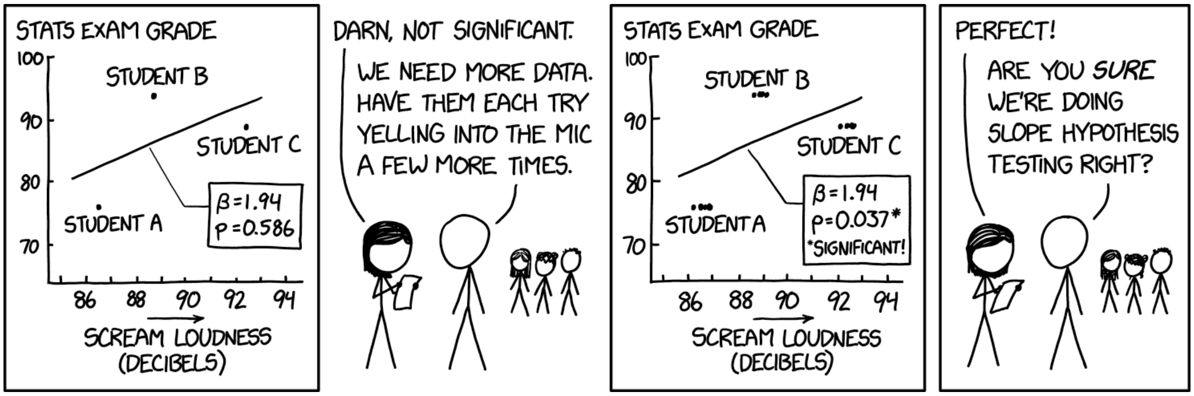
- The tea example is quite compelling, but in the real world hypothesis testing is highly misinterpreted
- Some journals are now banning significance based hypothesis testing, and the American Statistical Association has had to put out statements regarding the misuse of p values
- When we run a regression model, our hypothesis that we are rejecting is not that $\beta_1 = 0$ (if we reject that then it must be that $\beta_1 \neq 0$ which implies there is a causal relationship). Rather, we are rejecting the claim that $\beta_1 = 0$ **and all of our additional modeling assumptions are correct**

When Hypothesis testing goes wrong

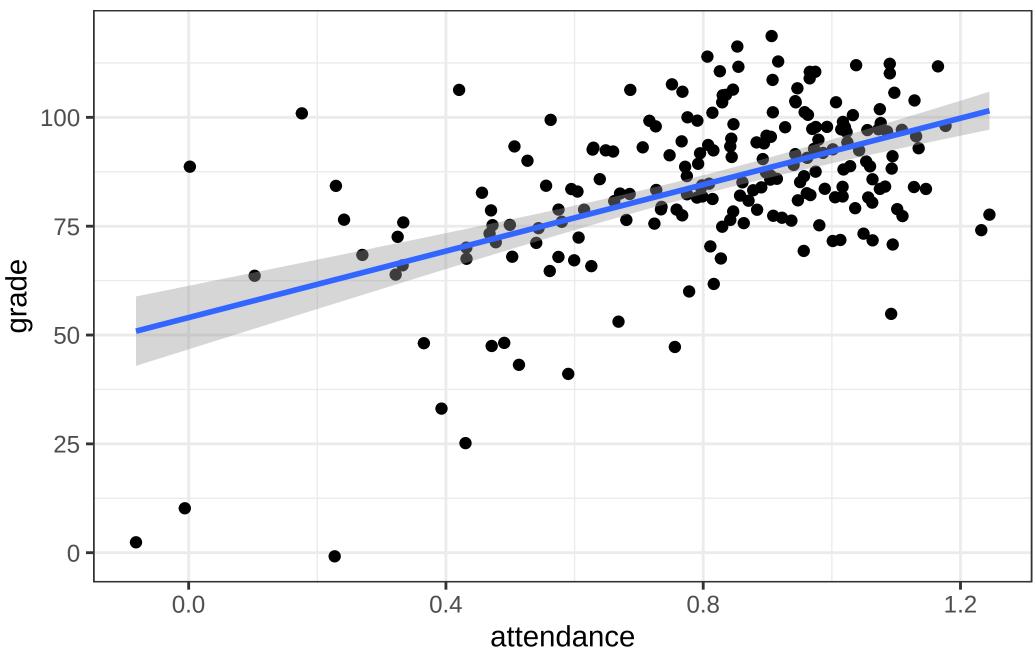
- In particular, we assumed in calculating the distribution of $\widehat{\beta}_1$ that this was an unbiased estimate (exogeneity), and we calculated the variance using a formula that assumed uncorrelated error terms.
- When we reject H_0 we are just saying that **at least one one of these assumptions is (statistically) incorrect**

- Note that if we have an extremely large sample size, even slight differences will be significant (e.g. $\beta_1 = .001 \neq 0$), but this means even very small modeling assumption errors will also result in significance. On big data you will almost always get p<.01, but that means virtually nothing

Cluster correlation



When Hypothesis testing goes wrong: real data

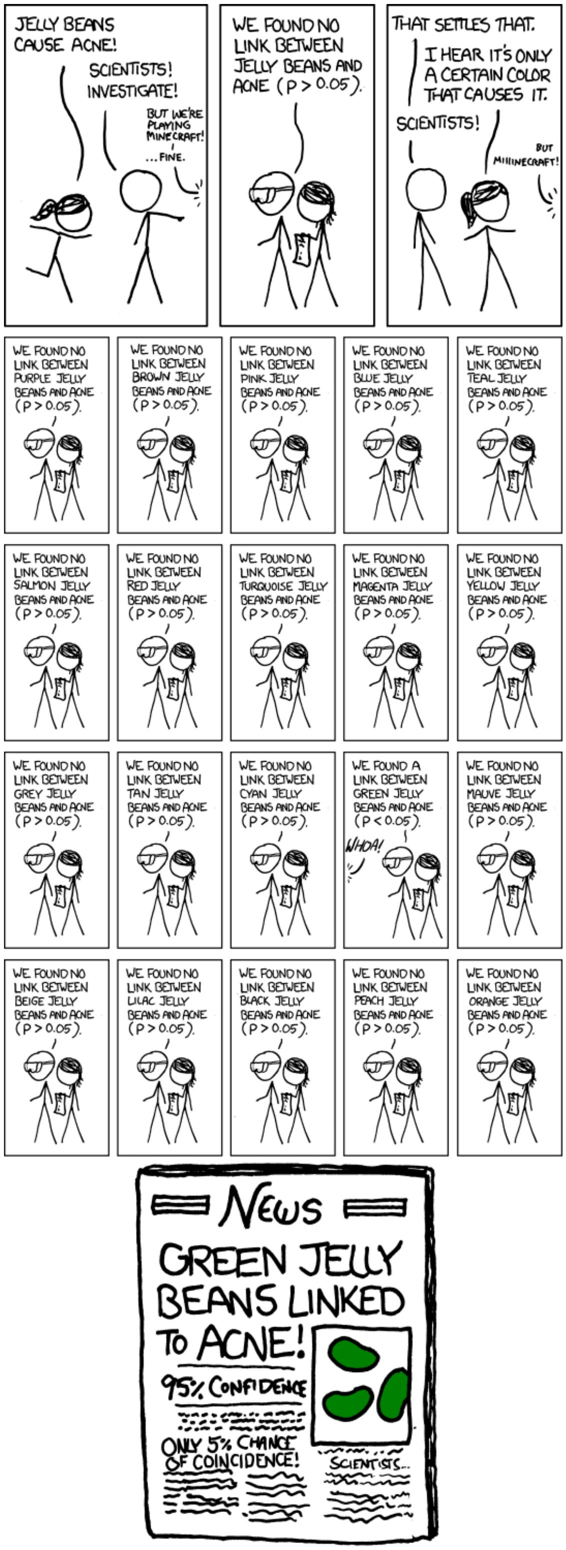


```
Call:
lm(formula = grade ~ attendance, data = dt)

Residuals:
    Min      1Q  Median      3Q     Max 
-63.514  -8.129   0.613   9.083  40.188 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   54.011      3.703  14.588  < 2e-16 ***
attendance    38.164      4.377   8.719 9.41e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.85 on 206 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.2695,    Adjusted R-squared:  0.266 
F-statistic: 76.02 on 1 and 206 DF,  p-value: 9.408e-16
```

When Hypothesis testing goes wrong: real data

- Our p value is 0.000000000000000941. Highly Significant!
- We reject H_0 that $\beta_1 = 0$, but we absolutely cannot claim that this is a causal effect because there is no exogeneity.
- Rather, we're concluding that **all of our assumptions combined** are implausible. i.e. it almost certainly cannot be the case that $\beta_1 = 0$ AND we have exogeneity in our model AND our error terms are uncorrelated AND our true model is linear.
- Note that this says almost nothing!

Yet another type of wrong



So is there any use for hypothesis testing

- First, you must show that your modeling assumptions are valid. For example, claim exogeneity by using a randomized control trial or other compelling quasi-experimental design
- Second, make conservative estimates in your other assumptions. e.g. overstating the variance in your model will be more convincing than understating your variance
- Third, interpret your results within the context of how precise your estimate is
- Significance testing is a small but important part of research

Hypothesis testing: Summary of steps

- Calculate $\hat{\beta}_1$ and $se(\hat{\beta}_1)(= \sigma_{\hat{\beta}_1})$ (this will be given to you as the output to a regression)
- Compute the z-score (the number of standard deviations from the mean) : $z = \frac{\hat{\beta}_1 - 0}{se(\hat{\beta}_1)}$
- Calculate $P(|Z| \geq z)$, where $Z \sim N(0,1)$
  - Look it up in a normal table, or use 1-(pnorm(z)-pnorm(-z)) in R
- Compare p to α. If $p < \alpha$, reject H_0, otherwise fail to reject H_0

Hypothesis testing: Summary of steps

- For a confidence interval, instead calculate $\hat{\beta}_1 \pm z_\alpha * se(\hat{\beta}_1)$
  - z_α is looked up via a table, or qnorm in R. It's 1.96 for 95% confidence