Advanced Regression Analysis

4. Inference and Interpretation of Linear Regression

GOVT 6029 - Spring 2020

Cornell University

Housekeeping

- Class time

Housekeeping

- Class time
    - Proposal: Thursday 10 am

Housekeeping

- Class time
    - Proposal: Thursday 10 am
- Assignments
    - Proposal:

Housekeeping

- Class time
  - Proposal: Thursday 10 am
- Assignments
  - Proposal:
    - Problem set 2 due Friday Apr 10

Housekeeping

- Class time
    - Proposal: Thursday 10 am
- Assignments
    - Proposal:
        - Problem set 2 due Friday Apr 10
        - Problem set 3 due Monday Apr 27

Housekeeping

- Class time
  - Proposal: Thursday 10 am
- Assignments
  - Proposal:
    - Problem set 2 due Friday Apr 10
    - Problem set 3 due Monday Apr 27
    - No problem set 4

Housekeeping

- Class time
    - Proposal: Thursday 10 am
- Assignments
    - Proposal:
        - Problem set 2 due Friday Apr 10
        - Problem set 3 due Monday Apr 27
        - No problem set 4
        - Poster and Paper due May 22

Housekeeping

- Class time
    - Proposal: Thursday 10 am
- Assignments
    - Proposal:
        - Problem set 2 due Friday Apr 10
        - Problem set 3 due Monday Apr 27
        - No problem set 4
        - Poster and Paper due May 22
        - Poster Session TBD Fall 2020

Housekeeping

- Class time
    - Proposal: Thursday 10 am
- Assignments
    - Proposal:
        - Problem set 2 due Friday Apr 10
        - Problem set 3 due Monday Apr 27
        - No problem set 4
        - Poster and Paper due May 22
        - Poster Session TBD Fall 2020

Statistical inference

Estimating the linear relationship between x and y is useful

But how certain are we of these estimates?

Statistical inference

Estimating the linear relationship between x and y is useful

But how certain are we of these estimates?

Even assuming

- linearity of the relationship between x and y,
- normality of $\mathrm{E}(y|x)$
- independence of each observation
- no correlation between ε and x

Are we sure we got the “right” estimate?

Statistical inference

Estimating the linear relationship between x and y is useful

But how certain are we of these estimates?

Even assuming

- linearity of the relationship between x and y,
- normality of $\mathrm{E}(y|x)$
- independence of each observation
- no correlation between ε and x

Are we sure we got the “right” estimate?

If we sampled more data, would we get the same estimates?

How close to the truth would our estimates be on average?

Statistical inference

If we sampled more data, would we obtain the same estimates?

Statistical inference

If we sampled more data, would we obtain the same estimates?

Usually, we can't answer this question with certainty

But we can **quantify our uncertainty** regarding $\hat{\beta}_0$, $\hat{\beta}_1$, etc.

Statistical inference

If we sampled more data, would we obtain the same estimates?

Usually, we can't answer this question with certainty

But we can **quantify our uncertainty** regarding $\hat{\beta}_0$, $\hat{\beta}_1$, etc.

Our goal as quantitative scientists:

Statistical inference

If we sampled more data, would we obtain the same estimates?

Usually, we can't answer this question with certainty

But we can **quantify our uncertainty** regarding $\hat{\beta}_0$, $\hat{\beta}_1$, etc.

Our goal as quantitative scientists:

Estimating unknowns and quantifying the uncertainty of those estimates

Estimates without measures of uncertainty aren't trustworthy or useful

Statistical inference

Where we are headed:

- Uncertainty of regression results

Statistical inference

Where we are headed:

- Uncertainty of regression results

Start simpler:

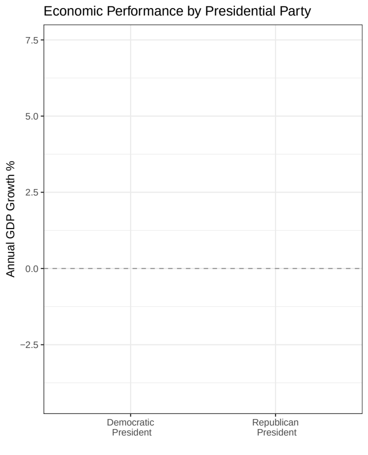
- Uncertainty of difference in two means

Difference between these means statistically meaningful?

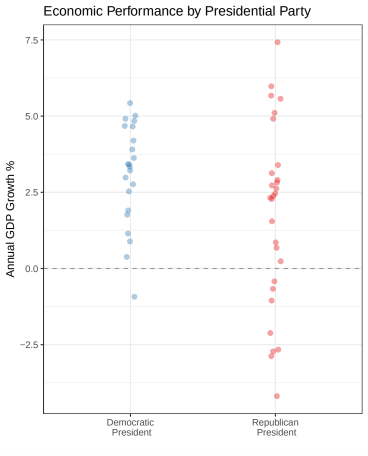
Example:

Do Democratic presidents enjoy statistically significantly higher growth?

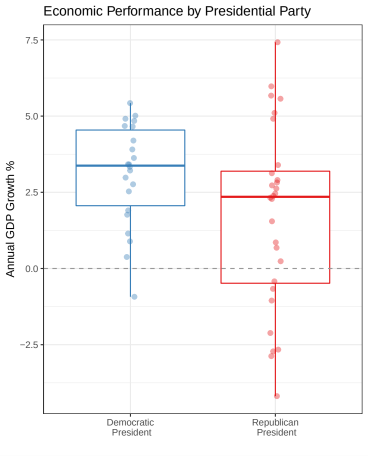
Difference between these means statistically meaningful?



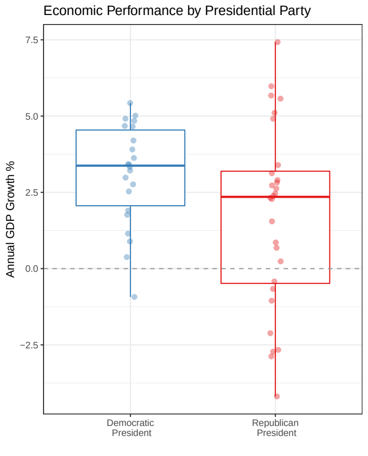
Difference between these means statistically meaningful?



Difference between these means statistically meaningful?



Difference between these means statistically meaningful?



Note difference between this question and the issue of substantive significance

Comparison of two means

How do we quantify certainty of estimates exactly?

Comparison of two means

How do we quantify certainty of estimates exactly?

Recall that to estimate the mean of a population from a sample we calculate

$$\mathrm{E}(x) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

It follows that the diff. of two means can be estimated by:

$$\mathrm{E}(x-y) \quad = \mathrm{E}(x) - \mathrm{E}(y)$$

Comparison of two means

How do we quantify certainty of estimates exactly?

Recall that to estimate the mean of a population from a sample we calculate

$$\mathrm{E}(x) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

It follows that the diff. of two means can be estimated by:

$$\mathrm{E}(x-y) \quad = \mathrm{E}(x) - \mathrm{E}(y) \quad = \frac{1}{n}\sum_{i=1}^{n} x_i - y_i$$

If n contained all the population, we'd be done

Comparison of two means

How do we quantify certainty of estimates exactly?

Recall that to estimate the mean of a population from a sample we calculate

$$\mathrm{E}(x) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

It follows that the diff. of two means can be estimated by:

$$\mathrm{E}(x-y) \quad = \mathrm{E}(x) - \mathrm{E}(y) \quad = \frac{1}{n}\sum_{i=1}^{n} x_i - y_i$$

If n contained all the population, we'd be done

If n is a sample, we can't be certain the sample is representative "enough" to produce the an accurate estimate of $\mathrm{E}(x-y)$

Comparison of two means

Suppose the sampling was done at random.

Comparison of two means

Suppose the sampling was done at random.

Then for each sample, we get a different $\mathrm{E}(\tilde{x} - \tilde{y})$

Comparison of two means

Suppose the sampling was done at random.

Then for each sample, we get a different $\mathrm{E}(\tilde{x} - \tilde{y})$

If we could describe the distribution of $\mathrm{E}(\tilde{x} - \tilde{y})$,

we could say something about how likely it is (in principle) that $\mathrm{E}(\tilde{x} - \tilde{y})$ takes on values in a certain range

Comparison of two means

Suppose the sampling was done at random.

Then for each sample, we get a different $\mathrm{E}(\tilde{x} - \tilde{y})$

If we could describe the distribution of $\mathrm{E}(\tilde{x} - \tilde{y})$,

we could say something about how likely it is (in principle) that $\mathrm{E}(\tilde{x} - \tilde{y})$ takes on values in a certain range

$\rightarrow$ how certain we are that our estimate of $\mathrm{E}(x - y)$ is right

Comparison of two means

Suppose the sampling was done at random.

Then for each sample, we get a different $\mathrm{E}(\tilde{x} - \tilde{y})$

If we could describe the distribution of $\mathrm{E}(\tilde{x} - \tilde{y})$,

we could say something about how likely it is (in principle) that $\mathrm{E}(\tilde{x} - \tilde{y})$ takes on values in a certain range

→ how certain we are that our estimate of $\mathrm{E}(x - y)$ is right

To get there, we need to introduce some new distributions and concepts

The χ^2 distribution

What if we have a variable that is the sum of n squared independent standard Normal RVs: $X^2 = x_1^2 + x_2^2 + \ldots x_n^2$

The χ^2 distribution

What if we have a variable that is the sum of n squared independent standard Normal RVs: $X^2 = x_1^2 + x_2^2 + \ldots x_n^2$

Sum of a finite set of Normal RVs, so the Normal only applies approximately

What distribution does this sum really follow?

The χ^2 distribution

What if we have a variable that is the sum of n squared independent standard Normal RVs: $X^2 = x_1^2 + x_2^2 + \ldots x_n^2$

Sum of a finite set of Normal RVs, so the Normal only applies approximately

What distribution does this sum really follow?

The χ^2 (chi-square) distribution,

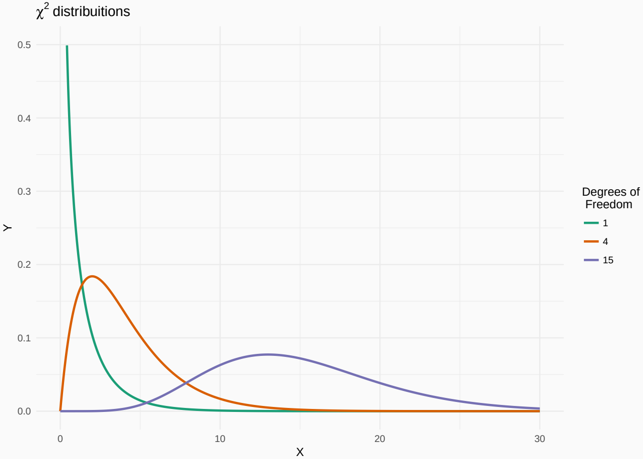
$$\chi^2(X_n^2) = \frac{1}{2^{n/2}\Gamma(n/2)}(x^2)^{(n-2)/2}\exp(-x/2)$$

which has "degrees of freedom" n

$\Gamma(\cdot)$ is the Gamma function, an interpolated factorial

$\mathrm{E}(\chi^2) = n$ and $\mathrm{Var}(\chi^2) = 2n$

χ^2 distributions



The t distribution

The χ_2 is a key building block for an even more useful distribution

The t distribution

The χ_2 is a key building block for an even more useful distribution

Suppose Z is standard normal and X^2 is distributed χ^2 with n degrees of freedom.

The t distribution

The χ_2 is a key building block for an even more useful distribution

Suppose Z is standard normal and X^2 is distributed χ^2 with n degrees of freedom.

Define

$$t = \frac{Z}{\sqrt{X^2/n}}$$

which is distributed t with n degrees of freedom:

$$f_t(t, n) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\Gamma(n/2)}} \times \frac{1}{(1 + t^2/n)^{(n+1)/2}}$$

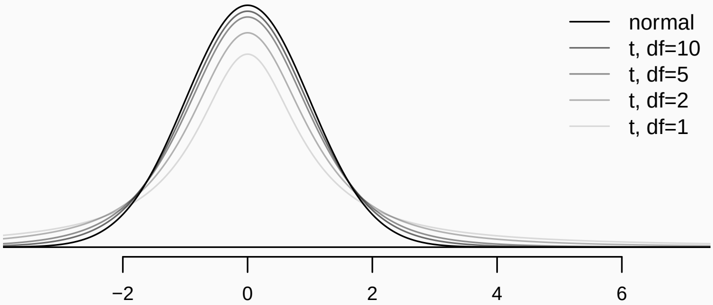
$\mathrm{E}(t) = 0$ (we could change this)

$\mathrm{Var}(t) = n/(n-2)$ for $n > 2$. Not defined for $n = 1$.

t distributions

As the degrees of freedom grow, the t distribution approximates the Normal

For low degrees of freedom, the t has fatter tails



The t distribution

Suppose we have a variable t that is t-distributed with mean 0 and 5 dfs

That is, $\mathrm{P}(t) = f_t(5)$

How large would t need to be for us to doubt it came from this distribution?

The t distribution

Suppose we have a variable t that is t-distributed with mean 0 and 5 dfs

That is, $\mathrm{P}(t) = f_t(5)$

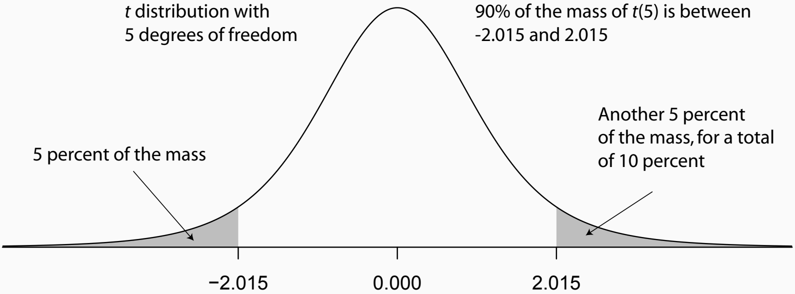
How large would t need to be for us to doubt it came from this distribution?

Put another way, what are the “critical” values of T we would see just

- once in 10 draws?
- once in 20 draws?
- once in 100 draws?

Put still another way,
which critical values will bound the 90% (or 95%, or 99%) most ordinary t draws?

Areas under the t

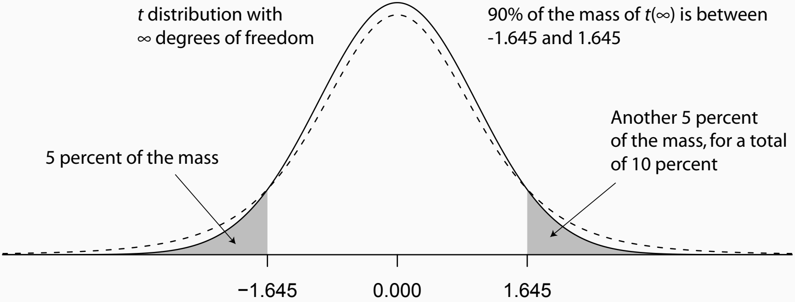


A unusual value is one in the tails. Critical values = cutoff for "unusualness"

To get the curve: `dt(x,df)`

To get critical values: `qt(quantile,df)`. Here, quantiles are 0.05 & 0.95

Areas under the t

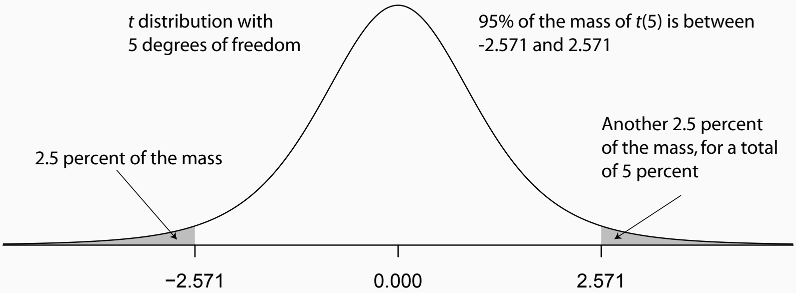


The degrees of freedom reflect how much information we have

More information makes the tails thinner

Critical values shrink; estimates get more certain

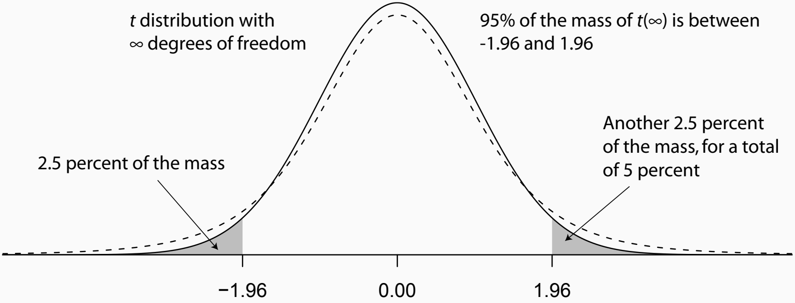
Areas under the t



Going back to the $df = 5$ case, notice we can choose what constitutes unusual

Here, we've raise the bar: only the 5% most extreme values are unusual

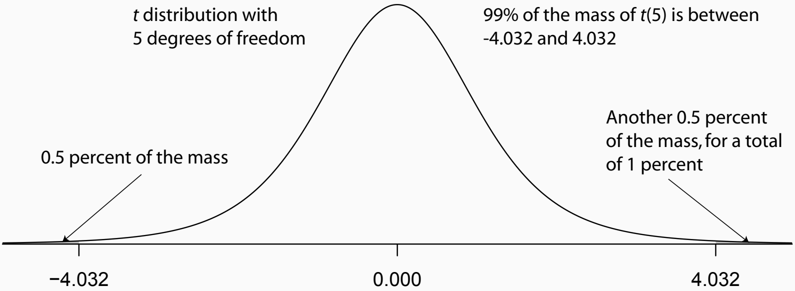
Areas under the t



The infinite degrees of freedom critical values for the 95% case

This is the most widely used standard for whether a result is unusual

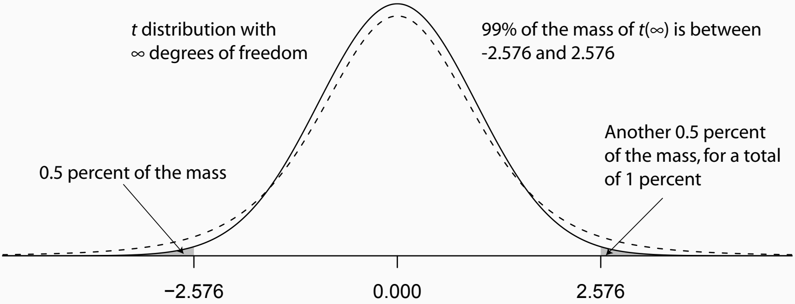
Areas under the t



The most stringent standard is 99%

In this case, a draw from the t must be in the 1% most extreme region to be considered unusual

Areas under the t



The infinite degrees of freedom case for 99%

Critical values of the t distribution

We can state how unusual an observation is under the assumption that it is $t(n)$

Test level	$\mathrm{df} = 5$	$\mathrm{df} = \infty$
0.1 level / 90%	2.015	1.645
0.05 level / 95%	2.571	1.960
0.01 level / 99%	4.032	2.576

These will be very useful for quantifying the uncertainty of estimates

Ways to summarize uncertainty

Now that we have some statistical foundations, what will we build?

There are two main ways to summarize uncertainty:

- **Confidence intervals**
- **Significance tests**

Ways to summarize uncertainty

Now that we have some statistical foundations, what will we build?

There are two main ways to summarize uncertainty:

- **Confidence intervals**
- **Significance tests**

I (and many other methodologists) prefer confidence intervals

- Somewhat easier to interpret
- Somewhat less likely to mislead (though not foolproof)

Ways to summarize uncertainty

Now that we have some statistical foundations, what will we build?

There are two main ways to summarize uncertainty:

- **Confidence intervals**
- **Significance tests**

I (and many other methodologists) prefer confidence intervals

- Somewhat easier to interpret
- Somewhat less likely to mislead (though not foolproof)

Significance tests were the standard approach for many decades & remain common

Ways to summarize uncertainty

Now that we have some statistical foundations, what will we build?

There are two main ways to summarize uncertainty:

- **Confidence intervals**
- **Significance tests**

I (and many other methodologists) prefer confidence intervals

- Somewhat easier to interpret
- Somewhat less likely to mislead (though not foolproof)

Significance tests were the standard approach for many decades & remain common

Need to know both

Significance tests

Suppose we wanted to compare the “unusualness” of two possible values of $\bar{x}$.

We might wonder:
if the population $\bar{x}$ is 0, how unlikely would it be to observe a sample $\bar{x} = 1.89$?

We can calculate this probability, under the assumption that $\bar{x}/\mathrm{se}(\bar{x})$ is distributed t.

$\mathrm{se}(\bar{x})$ is the standard error of the (estimated) mean

Ways to summarize uncertainty

We can calculate this probability, under the assumption that $\bar{x}/\mathrm{se}(\bar{x})$ is distributed t.

$\mathrm{se}(\bar{x})$ is the standard error of the (estimated) mean

That is,
how much we would expect to miss the truth by on average with repeated sampling

Ways to summarize uncertainty

We can calculate this probability, under the assumption that $\bar{x}/\mathrm{se}(\bar{x})$ is distributed t.

$\mathrm{se}(\bar{x})$ is the standard error of the (estimated) mean

That is,
how much we would expect to miss the truth by on average with repeated sampling

Why assume $\bar{x}/\mathrm{se}(\bar{x})$ is t distributed, and not, say, Normal?

Because the form the se takes is $\mathrm{se}(\bar{x}) = \sqrt{\sigma^2/n}$

This matches the denominator of $t = Z/\sqrt{X^2/N}$,
and only approximates the Normal for large N

Significance tests

Neyman-Pearson hypothesis testing in general:

- Set up a baseline, or **null** hypothesis for value of $\bar{x}$
- Calculate the prob of seeing a $\bar{x}$ as distant from the **null** as you did (this is the "p-value")
- If that probability is above a pre-committed threshold, "reject" the alternative hypothesis (your finding) in favor of the null

Notice the implicit notion that you could sample more data.

Frequentist inference: emphasis on data you could (but didn't) sample

One of two main branches of inference (the other is Bayesian)

Significance tests

Type I error: Probability of falsely rejecting the null

Type II error: Probability of falsely accepting the null

Significance tests minimize Type I error at the expense of Type II

Seen by some as "conservative"

My view: conservative is another way of saying "wrong in a certain direction",

Significance tests

Type I error: Probability of falsely rejecting the null

Type II error: Probability of falsely accepting the null

Significance tests minimize Type I error at the expense of Type II

Seen by some as "conservative"

My view: conservative is another way of saying "wrong in a certain direction", which is another way of saying biased.

Significance tests

Type I error: Probability of falsely rejecting the null

Type II error: Probability of falsely accepting the null

Significance tests minimize Type I error at the expense of Type II

Seen by some as “conservative”

My view: conservative is another way of saying “wrong in a certain direction”, which is another way of saying biased.

Who privileged the null over the alternative?

We chose both null and alternative hypotheses—
why should we want to treat them asymetrically?

Does it make sense to privilege your “second-best” hypothesis?

Significance tests

Type I error: Probability of falsely rejecting the null

Type II error: Probability of falsely accepting the null

Significance tests minimize Type I error at the expense of Type II

Seen by some as "conservative"

My view: conservative is another way of saying "wrong in a certain direction", which is another way of saying biased.

Who privileged the null over the alternative?

We chose both null and alternative hypotheses—
why should we want to treat them asymetrically?

Does it make sense to privilege your "second-best" hypothesis?

→ Reporting confidence intervals rather than p-values avoids null hypotheses and helps focus on the substance of your result

Significance test for a single mean

Generally, t-statistics are the ratio of the estimate to its standard error

In this case,

$$t = \frac{\bar{x}}{\mathrm{se}(\bar{x})}$$

The standard error of the mean is

$$\sigma_{\bar{x}} = \sqrt{\frac{\sigma_x^2}{n_x}} = \frac{\sigma_x}{\sqrt{n_x}}$$

Notice this gets smaller the more data we have

t can be compared to the critical value of t with $n - 1$ degrees of freedom for a significance level, α (e.g., $\alpha = 0.05$)

Or we can calculate the probability of getting so extreme a t from a random draw from the t distribution with $n - 1$ degrees of freedom

Significance tests

So far, we calculated t-stats for a mean, to test whether it is different from zero.

We can do this in `R` as follows:

Significance tests

So far, we calculated t-stats for a mean, to test whether it is different from zero.

We can do this in `R` as follows:

But we're really more interested in whether the Dem and Rep GDP means are significantly different

So we need a t-test for the difference in means.

t-test for comparison of means

As with a single mean, we will calculate a t-statistic:

$$t = \frac{\bar{x} - \bar{y}}{\mathrm{se}(\bar{x} - \bar{y})}$$

then check if the t-statistic exceeds the chosen critical value
or simply calculate the p of seeing so large a t

Because the two samples may have different sizes,
the form of the standard error here is a bit messy:

$$\mathrm{se}(\bar{x} - \bar{y}) = \sqrt{\left(\frac{(n_x - 1)\hat{\sigma}_x^2 + (n_y - 1)\hat{\sigma}_y^2}{n_x + n_y - 2}\right) \times \left(\frac{1}{n_x} + \frac{1}{n_y}\right)}$$

t-test for comparison of means

Unfortunately, the number of degrees of freedom, ν, is now ambiguous,
also because the samples could be different sizes

An estimate of the dfs for the comparison of means of different-sized samples is:

$$\hat{\nu} = \frac{\left(\frac{\hat{\sigma}_x^2}{n_x} + \frac{\hat{\sigma}_y^2}{n_y}\right)^2}{\frac{\hat{\sigma}_x^4}{n_x^2(n_x-1)} + \frac{\hat{\sigma}_y^4}{n_y^2(n_y-1)}}$$

(Don't worry, you'll never need to do this by hand)

t-tests for difference of means in R

It’s easy to do this test in R

t-tests for difference of means in `R`

It's easy to do this test in `R`

How do we interpret the above, in plain English?

Confidence intervals

Some notation:

$\bar{x}$	The mean of a sample of n values of x
$\bar{x}^{\text{population}}$	The true mean from the population of x; unknown
$\bar{x}^{\text{upper}}$	Upper bound of a confidence interval around $\bar{x}$
$\bar{x}^{\text{lower}}$	Lower bound of a confidence interval around $\bar{x}$

Confidence intervals

What we'd really like to know:

The (objective) probability that the population mean lies between $\bar{x}^{\text{lower}}$ & $\bar{x}^{\text{upper}}$

We can never know this from an incomplete sample.

No one will ever devise a technique to show us.

Not possible within frequentist inference.

Confidence intervals

This is impossible to find:

The (objective) probability that the population mean lies between $\bar{x}^{\text{lower}}$ & $\bar{x}^{\text{upper}}$

Now what?

Confidence intervals: Option 1

Drop the word “objective”, and calculate a subjective probability. This is the method used by Bayesian inference:

- Based on an initial, subjective assessment (e.g., personal reading of past research), define a prior distribution of $\bar{x}^{\text{population}}$, $P(\bar{x}^{\text{population}})$.
- One way to read this prior distribution:
  A set of intervals within which you believe $\bar{x}^{\text{population}}$ lies with $1-\alpha$ probability.
- Thus we already have (by assumption alone) an answer to our question. But we want to update that subjective probability interval to account for inference from our data.
- We update our distibution of $\bar{x}^{\text{population}}$ by Bayes rule:
- $$P(\bar{x}^{\text{population}}|\bar{x}) = P(\bar{x}^{\text{population}})\frac{P(\bar{x}|\bar{x}^{\text{population}})}{P(\bar{x})}$$

Now we have a new set of $1-\alpha$ Bayesian credible intervals implied by $P(\bar{x}^{\text{population}}|\bar{x})$.

These intervals give the probability $\bar{x}^{\text{population}}$ is between $\bar{x}^{\text{lower}}$ and $\bar{x}^{\text{upper}}$

Confidence intervals: Option 2

If we want to remain "objective"—that is, ignore our prior beliefs—we cannot calculate a probability that $\bar{x}^{\text{population}}$ lies between $\bar{x}^{\text{lower}}$ and $\bar{x}^{\text{upper}}$.

The frequentist solution:

A new concept, called **confidence**, used in place of probability.

We are 95% **confident** that $\bar{x}^{\text{population}}$ lies between $\bar{x}^{\text{lower}}$ and $\bar{x}^{\text{upper}}$ when in repeated samples from the population, 95% of intervals constructed in such a fashion contain the truth $\bar{x}^{\text{population}}$

Why isn't this a probability?

It's a statement about the asymptotic properties of the t distribution, not the data we observed.

We just don't know, based on the draw we made, whether the truth lies inside the frequentist confidence interval. And we don't know the probability this is true

Confidence intervals for means

To get the $100(1-\alpha)$% confidence interval for the mean of x,

$$\begin{aligned} \bar{x}^{\text{lower}} &= \bar{x} - t_{\alpha/2,n-2}\hat{\sigma}_x \\ \bar{x}^{\text{upper}} &= \bar{x} + t_{\alpha/2,n-2}\hat{\sigma}_x \end{aligned}$$

where $t_{\alpha/2,n-2}$ is the critical value of the t distribution with $n-2$ degrees of freedom and a probability of $a/2$ to the right

Confidence intervals for means

To get the $100(1-\alpha)$% confidence interval for the mean of x,

$$\begin{aligned}\bar{x}^{\text{lower}} &= \bar{x} - t_{\alpha/2,n-2}\hat{\sigma}_x \\ \bar{x}^{\text{upper}} &= \bar{x} + t_{\alpha/2,n-2}\hat{\sigma}_x\end{aligned}$$

where $t_{\alpha/2,n-2}$ is the critical value of the t distribution with $n-2$ degrees of freedom and a probability of $a/2$ to the right

"We're 95% confident $\bar{x}^{\text{true}}$ is equal to $\bar{x}$, $\pm$ a few standard deviations of x"

- Plus or minus how many std devs? $t_{\alpha/2,n-2}$ many.
- For a large dataset, the 95% CI will tend to be $\pm 1.96\hat{\sigma}$
- For a large dataset, the 99% CI will tend to be $\pm 2.576\hat{\sigma}$

But you need to calculate the exact value using `qt(alpha/2,n-2)`

Confidence intervals for difference of means

To get the $100(1-\alpha)$% confidence interval for a difference of means,

$$\bar{x} - \bar{y} \pm t_{\alpha/2,\hat{\nu}} \sqrt{\frac{\hat{\sigma}_x^2}{n_x} + \frac{\hat{\sigma}_y^2}{n_y}}$$

where $t_{\alpha/2,\hat{\nu}}$ critical value of the t distribution with $\hat{\nu}$ degrees of freedom and a probability of $a/2$ to the right

$\hat{\nu}$ is estimated as before

Just leave this one up to `R`...

Confidence intervals

How do we report a confidence interval?

Democratic presidents enjoyed growth rates 1.37 points higher [95% CI: 0.01 to 2.72] than their Republican counterparts.

or

Democrats enjoyed 1.37 points higher growth than Republicans, with a 95 percent confidence interval of 0.01 to 2.72.

We could calculate any CI we wish: 90 percent, 80 percent, 50 percent, etc.

The most commonly used are: 90, 95, and 99.

Can we get 100 percent CIs?

Confidence intervals

How do we report a confidence interval?

Democratic presidents enjoyed growth rates 1.37 points higher [95% CI: 0.01 to 2.72] than their Republican counterparts.

or

Democrats enjoyed 1.37 points higher growth than Republicans, with a 95 percent confidence interval of 0.01 to 2.72.

We could calculate any CI we wish: 90 percent, 80 percent, 50 percent, etc.

The most commonly used are: 90, 95, and 99.

Can we get 100 percent CIs?

Not unless we can logically reject values outside that interval

Significance tests vs confidence intervals

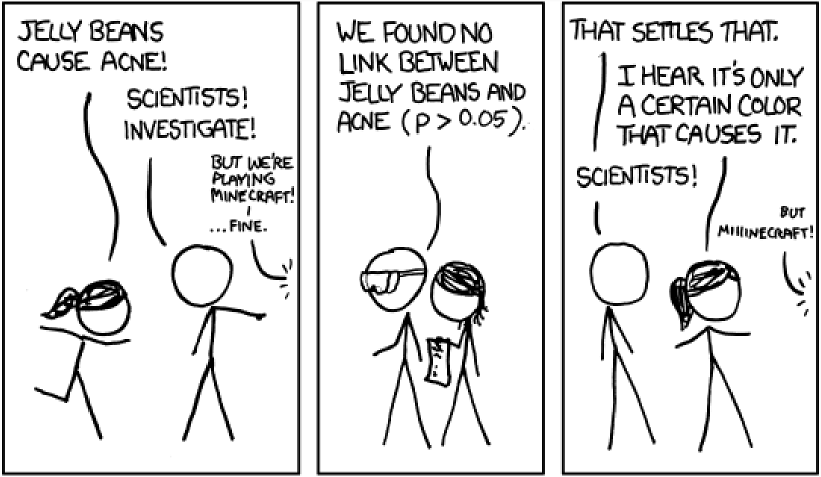
Problems with significance tests & confidence intervals:

- Simply a commitment to a certain error rate, given all assumptions met

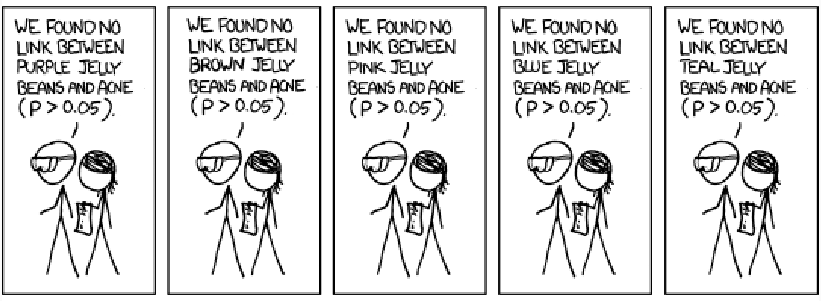
Significance tests vs confidence intervals

Problems with significance tests & confidence intervals:

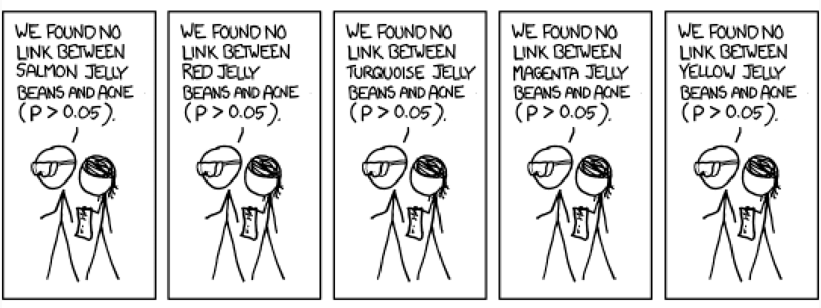
- Simply a commitment to a certain error rate, given all assumptions met
- Suppose null is really true. But eventually, someone will find a "significant" difference from any null by chance. If you publish all (and only) significant results, journals will be full of falsehoods, and file drawers of rejected papers full of truth.


JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
...FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MIIIINECRAFT!

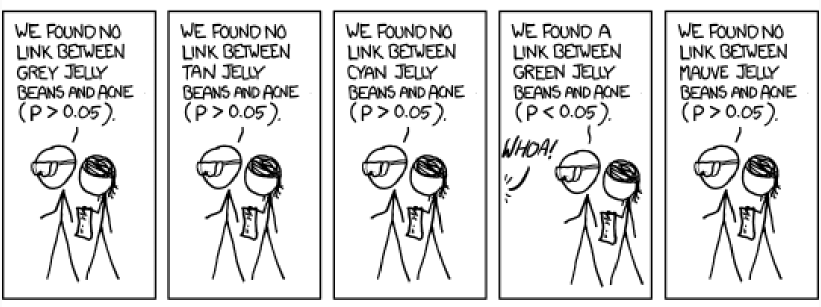
xkcd.com/882


WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).

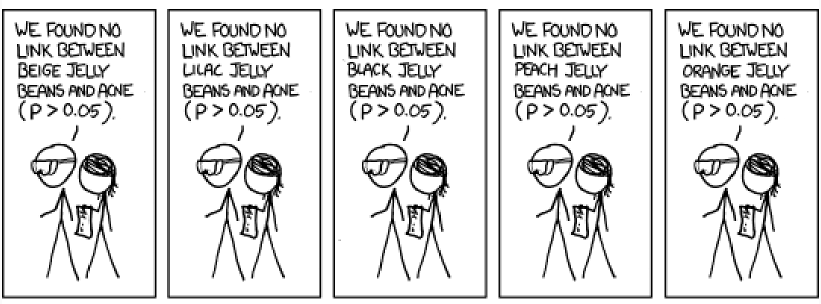
xkcd.com/882


WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).

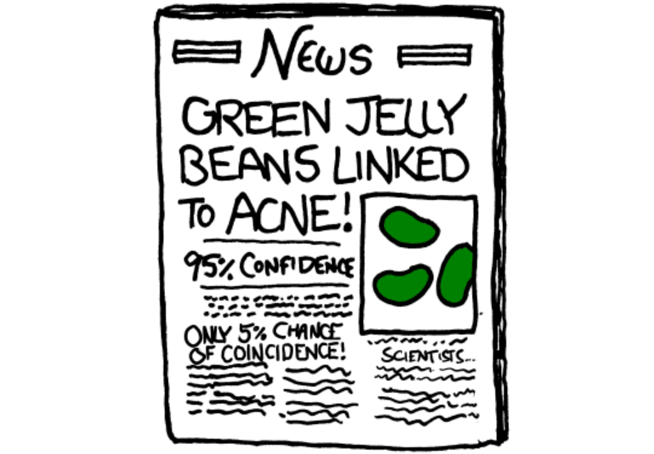
xkcd.com/882


WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).

xkcd.com/882


WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).

xkcd.com/882


NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

xkcd.com/882

Significance tests vs confidence intervals

Problems with significance tests & confidence intervals:

- Simply a commitment to a certain error rate, given all assumptions met
- Suppose null is really true. But eventually, someone will find a “significant” difference from any null by chance. If you publish all (and only) significant results, journals will be full of falsehoods, and file drawers of rejected papers full of truth.

Significance tests vs confidence intervals

Problems with significance tests that CIs overcome:

- Hypothesis tests are “weak” in the sense that they don’t tell you if the alternative hypothesis (your finding) was different from the new null hypothesis $X^{\text{null}} + 0.000001$

Significance tests vs confidence intervals

Problems with significance tests that CIs overcome:

- Hypothesis tests are “weak” in the sense that they don’t tell you if the alternative hypothesis (your finding) was different from the new null hypothesis $X^{\text{null}} + 0.000001$
- Encourage a focus on statistical significance rather than substantive significance. Star gazing, or “measuring” effects by p-values.

Significance tests vs confidence intervals

Confidence intervals are not perfect

Share many of the same limits & awkwardness of significance tests

And people tend to mistake them for probability intervals

Significance tests vs confidence intervals

Confidence intervals are not perfect

Share many of the same limits & awkwardness of significance tests

And people tend to mistake them for probability intervals

But which would you rather say or read:

- *Compared to Republicans, the effect of Democratic presidents on the economy is significantly positive at the 0.05 level.*

Significance tests vs confidence intervals

Confidence intervals are not perfect

Share many of the same limits & awkwardness of significance tests

And people tend to mistake them for probability intervals

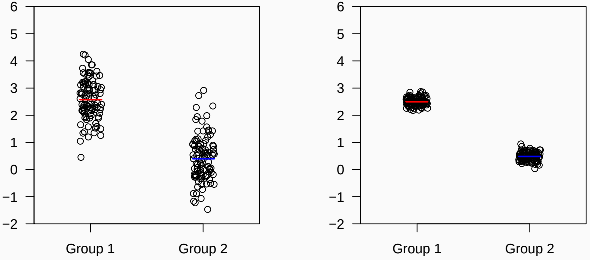
But which would you rather say or read:

- *Compared to Republicans, the effect of Democratic presidents on the economy is significantly positive at the 0.05 level.*
- Democratic presidents enjoyed 1.37 points higher growth than Republicans, with a 95 percent confidence interval of 0.01 to 2.72.

Even with the ambiguity of the words “significance” and “confidence”, the latter says more and says it more clearly, and avoids conflation of statistical and substantive significance

“These means are different.” Equally confident?

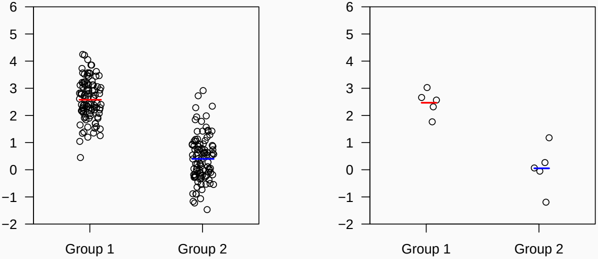
Same means in each pair; different variances:



$\mathcal{N}(2.5, 0.72)$, $\mathcal{N}(0.5, 0.72)$ $\mathcal{N}(2.5, 0.02)$, $\mathcal{N}(0.5, 0.02)$

Lower variance in this case gives more info per observation

“These means are different.” Equally confident?



Small samples lower significance. But may look like very strong relationships. $\mathcal{N}(2.5, 0.72)$ $\mathcal{N}(0.5, 0.72)$

Empirically, journals seem overly optimistic of statistical significance of small N results, and are likely to publish them.

Inference for regression

We can calculate confidence intervals and significance tests for any estimate, including regression coefficients

Regression coefficients are random variables

What is their sampling distribution?

$\rightarrow$ Distribution of $\hat{\beta}_1$ given repeated sampling of x, y from the population

Inference for regression

We can calculate confidence intervals and significance tests for any estimate, including regression coefficients

Regression coefficients are random variables

What is their sampling distribution?

→ Distribution of $\hat{\beta}_1$ given repeated sampling of x, y from the population

In practice, we sample x, y once, and then run the regression.

But imagine we did in thousands of times.

We'd pile up the estimated $\hat{\beta}_1$, and get a histogram.

What is the mean and variance of this distribution?

Inference for regression

What is the mean and variance of the dist of $\hat{\beta}_1$?

The mean is the expected value of $\hat{\beta}_1$.

If:

- the relationship between y and x really is linear,
- and x and ε really are independent,

then we expect $\hat{\beta}_1$ to match the truth, on average:

$$\mathrm{E}(\hat{\beta}_1) = \beta_1$$

Inference for regression

If:

- the relationship between y and x really is linear,
- and x and ε really are independent,
- and y really is Normal conditional on x,

the variance of $\hat{\beta}_1$ will be

$$\text{var}(\hat{\beta}_1) = \frac{\sigma_\varepsilon^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

Upshot: Can estimate the variance of $\hat{\beta}_1$ with a single draw from its distribution.
This is nothing short of magical

Inference for regression

To derive the variance-covariance matrix of the parameters $\boldsymbol{\beta}$:

$$\text{Var}(\hat{\boldsymbol{\beta}}) = \hat{\Sigma} = \text{E}\left[\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)'\right]$$

Note that:

$$\begin{aligned}
\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)' &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} - \boldsymbol{\beta} \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\left(\mathbf{X}\boldsymbol{\beta} - \boldsymbol{\varepsilon}\right) - \boldsymbol{\beta} \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon} - \boldsymbol{\beta} \\
&= \boldsymbol{\beta} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon} - \boldsymbol{\beta} \\
&= -(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}
\end{aligned}$$

Inference for regression

Substituting, we find

$$\begin{aligned}
\mathrm{Var}(\hat{\boldsymbol\beta}) = \hat{\Sigma} &= \mathrm{E}\left[\left(\hat{\boldsymbol\beta} - \boldsymbol\beta\right)\left(\hat{\boldsymbol\beta} - \boldsymbol\beta\right)'\right] \\
&= \mathrm{E}\left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol\varepsilon\boldsymbol\varepsilon'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\right] \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathrm{E}(\boldsymbol\varepsilon\boldsymbol\varepsilon')\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\sigma^2\mathbf{I})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\
&= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}
\end{aligned}$$

This form is analogous to the bivariate version above, with two differences:

It calculates every standard error at once

It calculates covariances, which allow the estimates $\hat{\beta}_i$ and $\hat{\beta}_j$ to be correlated

Inference for regression

Putting these together, and assuming $\mathrm{E}(y|x)$ really is Normal, we have the distribution of $\hat{\beta}_1$

$$\hat{\beta}_1 \sim \mathcal{N}\left(\beta_1, \frac{\sigma_\varepsilon^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right)$$

or, in matrix form, the joint distribution of all the $\hat{\boldsymbol{\beta}}$

$$\hat{\boldsymbol{\beta}} \sim \mathcal{MVN}\left(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\right)$$

Even if $\mathrm{E}(y|x)$ isn't Normal, this holds approximately as $n \to \infty$

NB: $\mathcal{MVN}(\cdot, \cdot)$ represents the Multivariate Normal distribution, which allows several variables to be both Normal and potentially correlated with each other

Inference for regression

We now have an estimate of the standard deviation of $\hat{\beta}_1$.

This is called the standard error of β_1:

$$\text{se}(\hat{\beta}_1) = \hat{\sigma}_{\hat{\beta}_1} = \sqrt{\frac{\hat{\sigma}^2_\varepsilon}{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

This is the second column of regression output in `R`.

The se's can also be found by taking the square roots of the diagonal elements of the variance-covariance matrix:

$$\text{se}(\hat{\boldsymbol{\beta}}) = \sqrt{\text{diag}\left(\sigma^2(\mathbf{X}'\mathbf{X})^{-1}\right)}$$

Notice that standard errors get bigger (less precise) when the data get nosier, or when the x's have little variation

Inference for regression

Assuming the β's are normally distributed, and their standard errors are χ^2 distributed, we can construct the following t-distributed test statistic

$$t = \frac{\hat{\beta}_1 - \beta_1^{\text{null}}}{\text{se}(\beta_1)}$$

commonly known as the t-statistic.

It is distributed t with $n - k - 1$ degrees of freedom

To conduct significance tests, just calculate the p-value

That is, the area under the tails beyond $(\hat{\beta}_1 - \beta_1^{\text{null}})/\text{se}(\beta_1)$ of a t-distribution with $n - k - 1$ dfs

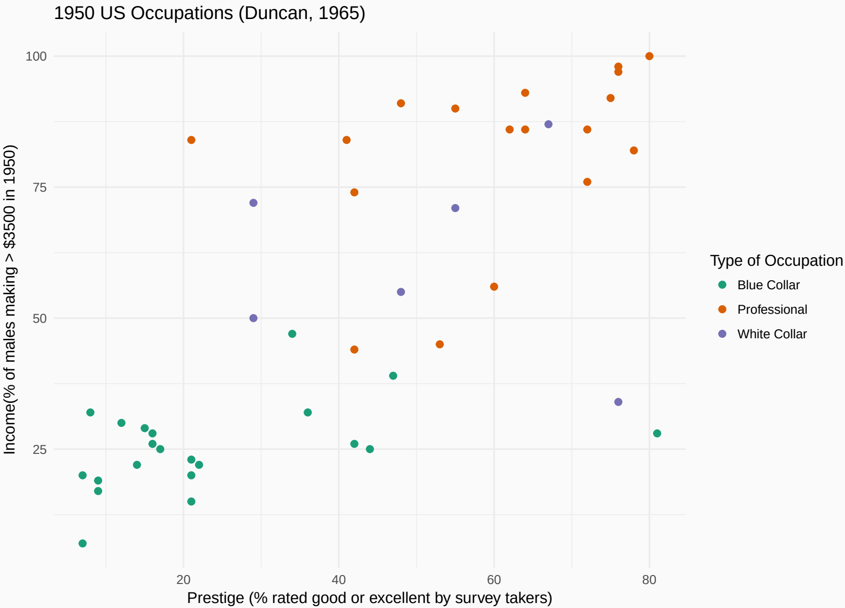
Example: Occupational Prestige & Income

Classic data from sociology. Three variables

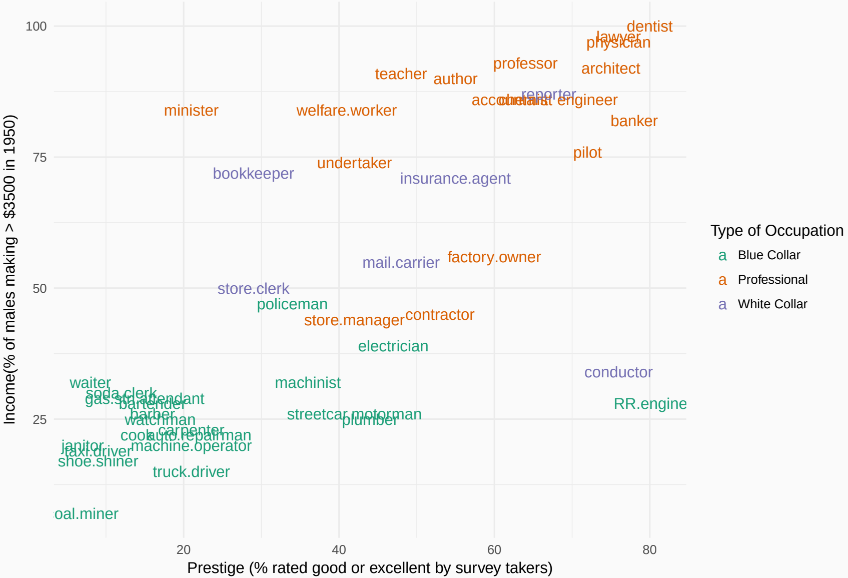
- Prestige of occupations, as rated by surveys
- Income of occupations (averaged across males)
- Type of occupation (blue collar, white collar, professional)

Data is in `R`.

Load the `car` library and run `data(Duncan)` and `help(Duncan)`


1950 US Occupations (Duncan, 1965)
Income(% of males making > $3500 in 1950)
100
75
50
25
20
40
60
80
Prestige (% rated good or excellent by survey takers)
Type of Occupation
● Blue Collar
● Professional
● White Collar

1950 US Occupations (Duncan, 1965)



Let's take this to R

Confidence intervals for regression coefficients

Standard errors, t-tests, and p-values take expertise to read

They are also subject to misinterpretation

(E.g., smaller p-values do not imply a bigger substantive effect)

CIs turn the standard errors into something more people can understand

To get the $100(1-\alpha)$% confidence interval for $\hat{\beta}_1$,

$$\begin{aligned}\hat{\beta}_1^{\text{lower}} &= \hat{\beta}_1 - t_{\alpha/2,n-k-1}\hat{\sigma}_{\hat{\beta}_1} \\ \hat{\beta}_1^{\text{upper}} &= \hat{\beta}_1 + t_{\alpha/2,n-k-1}\hat{\sigma}_{\hat{\beta}_1}\end{aligned}$$

Confidence intervals for regression coefficients

How to calculate CIs for coefficients in R

By hand: lower.95 ¡- betas - qt(0.025,42)*ses upper.95 ¡- betas + qt(0.025,42)*ses

Why are we using qt? Why 0.025?

The easy way: confint(lm.out,level=0.95)

2.5 (Intercept) -14.6857892 2.5564634 education 0.3475521 0.7441158 income 0.3572343 0.8402313

Confidence intervals for regression coefficients

Using confidence intervals, we can improve the initial summary table:

Table 1: Determinants of occupational prestige. Entries are linear regression parameters and their 95 percent confidence intervals.

		95% Conf Interval	
Variable	Estimate	Lower	Upper
Income	0.60	[0.36,	0.84]
Education	0.55	[0.38,	0.74]
Intercept	-6.06	[-14.69,	2.46]
N	45		
s.e.r.	13.4	(this is $\hat{\sigma}_{\varepsilon}$)	
R^2	0.83	(this line optional)	

Think about everything you put in these tables:

Confidence intervals for regression coefficients

Substantive & statistical significance

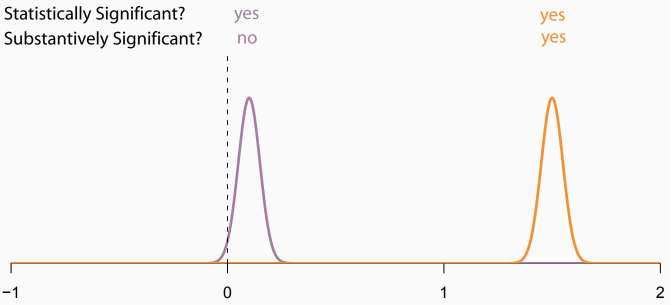
Don't over interpret p-values

They only show statistical significance

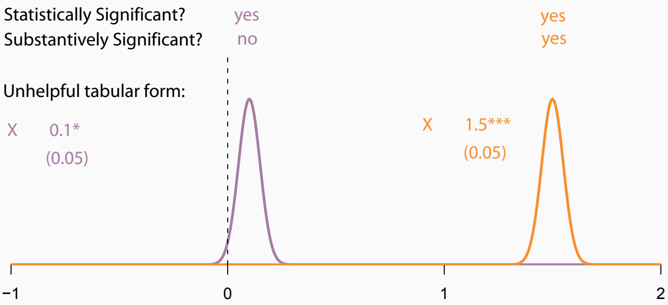
Statistical and substantive significance can interact

A look at some hypothetical distributions of $\hat{\beta}_1$ helps frame the possibilities

Perils of stargazing



Perils of stargazing

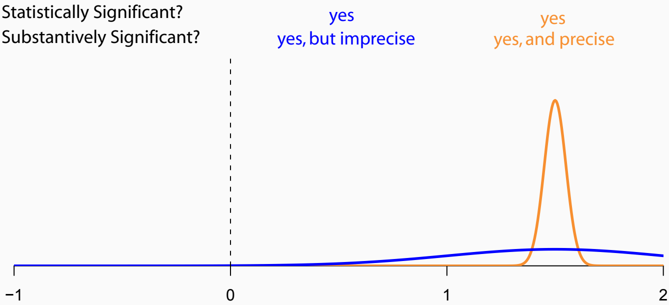


These estimated β's will both be starred in regression output.

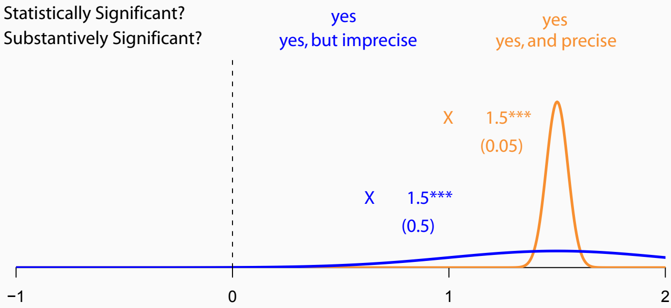
Often, only the estimate to the right will be significant in a substantive sense

The estimate on the left is a precise zero

Perils of stargazing



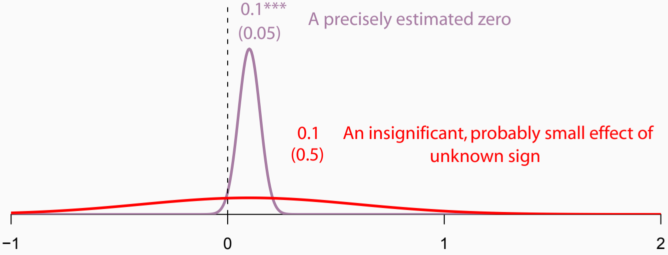
Perils of stargazing



These estimated β's will both be heavily starred in regression output.

They are both substantively significant as well, with identical point estimates

Perils of stargazing



How do you verify a null effect? Precise zeros

Sometimes, researchers mistake the precise zero for a positive effect

Confidence interval for expected values

We can calculate the CIs around $\hat{Y}$ as well.

For example, what is the 95% CI around $\widehat{\text{Prestige}}_c$ in:

$$\widehat{\text{Prestige}}_c = \hat{\beta}_0 + \hat{\beta}_1 \text{Income}_c + \hat{\beta}_2 \text{Education}_c$$

The uncertainty in each estimate will “combine” to form the uncertainty in $\widehat{\text{Prestige}}_c$.

In this example,

Confidence interval for expected values

How do we calculate confidence intervals around $\hat{y}$ in `R`?

1. Estimate the model
2. Choose hypothetical values of the covariate at which you want to calculate $\hat{y}$ and it's CI.
3. Use the `predict()` function to obtain the expected y and it's CI

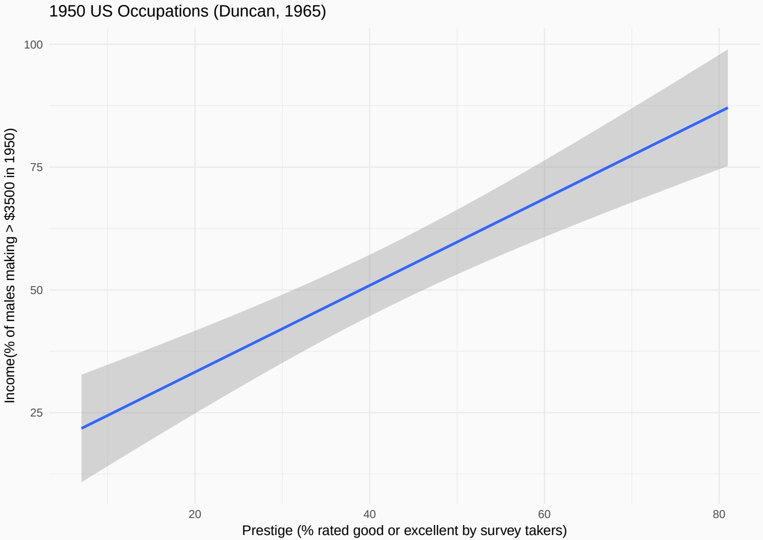
Some examples:

Confidence interval for expected values

The we code we run is very useful for adding confidence intervals to a plot.

We can run through a sequence of possible x values, holding z constant,
and predict y and it's confidence interval,
then plot the confidence interval as an envolpe around y

Confidence interval for expected values



Confidence interval for expected values

Interpretation:
All we can say with 95 percent confidence is that the line
—the relation b/w prestige and income–
lies in this envelope

Very useful to show, especially if the relationship is curved in some way

I prefer shaded regions to dotted lines. (lots of lines gets confusing)

You can make shaded regions using the `polygon()` command

Just be sure to plot the polygon before you add any points or lines, so it shows up behind them

Prosecutor's Fallacy

Examples sharpen distinction between credible and confidence intervals:

- You win the lottery. A prosecutor charges you with fraud, simply on the grounds that for any given person, winning the lottery without cheating is very unlikely.
- You are charged with a crime based on a DNA match. After searching through a database of 200,000 people's DNA, the prosecutor argues that you are probably guilty because a 99.999% confidence interval contains your DNA and DNA found at the scene of the crime.

Is the statistical reasoning here sound? What if these were credible intervals?

Wait! What are degrees of freedom (df)?

Degrees of freedom:
The number of separate pieces of information used to calculate a statistic

“separate” $=$ “freely movable”

Not the same thing as the number of observations (may be the same as N or less)

Relevance: how many quantities could we estimate from a set of data?

What are degrees of freedom (df)?

How many separate pieces of unspecified information to estimate?

$\exists$ two numbers, x_1 and x_2 2 pieces

What are degrees of freedom (df)?

How many separate pieces of unspecified information to estimate?

$\exists$ two numbers, x_1 and x_2 2 pieces

Now I decide to add an assumption regarding the value of x_2

$\exists$ two numbers, x_1 and 3 1 piece

What are degrees of freedom (df)?

How many separate pieces of unspecified information to estimate?

$\exists$ two numbers, x_1 and x_2 2 pieces

Now I decide to add an assumption regarding the value of x_2

$\exists$ two numbers, x_1 and 3 1 piece

Instead, suppose I assume I know the mean?

$\exists$ two numbers, x_1 and x_2, and $\bar{x} = 2$ 1 piece

What are degrees of freedom (df)?

How many separate pieces of unspecified information to estimate?

$\exists$ two numbers, x_1 and x_2 2 pieces

Now I decide to add an assumption regarding the value of x_2

$\exists$ two numbers, x_1 and 3 1 piece

Instead, suppose I assume I know the mean?

$\exists$ two numbers, x_1 and x_2, and $\bar{x} = 2$ 1 piece

How does this work at larger scales?

$\exists$ fifty numbers, $x_1, \ldots x_{50}$, and $\bar{x} = 2$ 49 pieces

What are degrees of freedom (df)?

How many separate pieces of unspecified information to estimate?

$\exists$ two numbers, x_1 and x_2 2 pieces

Now I decide to add an assumption regarding the value of x_2

$\exists$ two numbers, x_1 and 3 1 piece

Instead, suppose I assume I know the mean?

$\exists$ two numbers, x_1 and x_2, and $\bar{x} = 2$ 1 piece

How does this work at larger scales?

$\exists$ fifty numbers, $x_1, \ldots x_{50}$, and $\bar{x} = 2$ 49 pieces

$\exists$ fifty numbers, $x_1, \ldots x_{50}$, $\bar{x} = 2$, and $\sigma^2 = 0.5$ 48 pieces

What are degrees of freedom (df)?

Degrees of freedom (df): the remaining allowed ways you could move the data

If we make as many assumptions as there are observations, nothing left to estimate