The stats you need to know

Review of basic statistical concepts

We’ll touch on the basics

- Parameters and estimates
  - Confidence intervals for estimates
- Hypothesis testing about relationships between variables
  - Two numeric variables → regression
  - One numeric variable and one categorical → ANOVA

So you want to know the density of poppies in this field...

Density = (number of poppies)/(area)

If we could count every flower, and measure the area of the field, we could calculate the true density

But, we can't – we have to **estimate** the density from a **sample** instead

Problem with a sample – individual plots are variable

Each square is a 1 m^2 plot – number of poppies varies a lot among the plots

How do we get an estimate of the density of poppies per square meter from these?

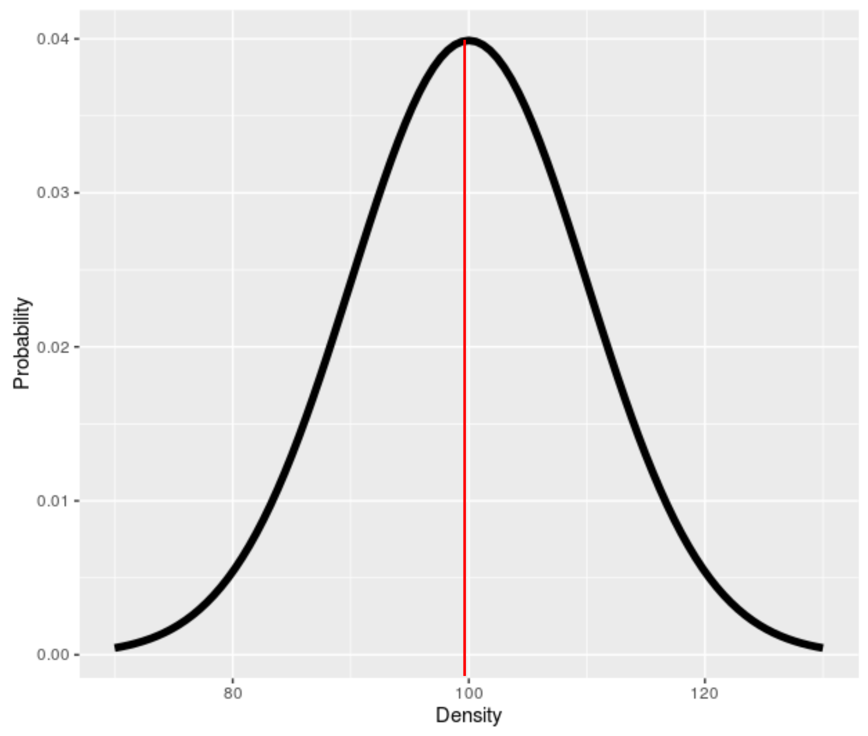
Solution: treat poppy density as a random variable

- Random variables are mathematical abstractions – models of variables subject to random variation
  - Random variables take different values → two repeated observations may give different results
  - Can't know the value of a random variable in advance of observing it
- We can use a mathematical model of the random variable to understand it better
  - Can make probability statements about what the random variable's value will be when it's observed
  - Example: the normal distribution as a model of poppy density

Distribution of densities of poppies in $1m^2$ plots

Individual plots have different numbers

If we measured every square meter, the mean would be the true mean density, which we call the population parameter: $\mu = 100$



$\mu = 100$

But, we don’t have complete information

- Usually, we don’t know μ, and all the information we have about it comes from a sample of data
- If we have a sample of 9 plots, with counts of:

  95, 68, 107, 93, 101, 113, 107, 100, 106

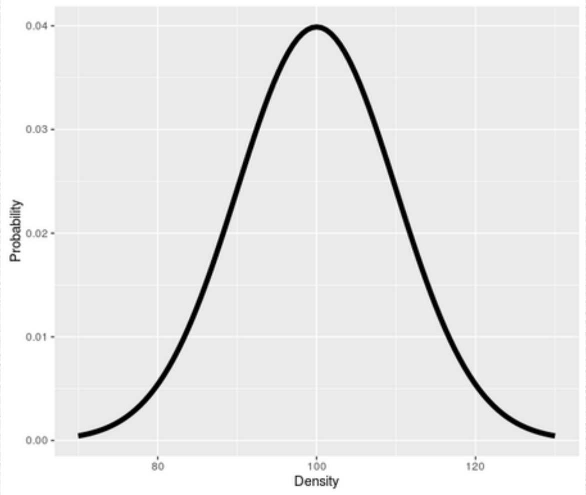
  the mean of this sample ($\bar{x}$) is 98.89
- We we can treat this sample mean ($\bar{x}$) as an **estimate** of the population parameter (μ)
- The amount of variation among the data values is measured with the standard deviation (s), which is 13.18
- How good an estimate of μ is $\bar{x}$?

Sampling variation

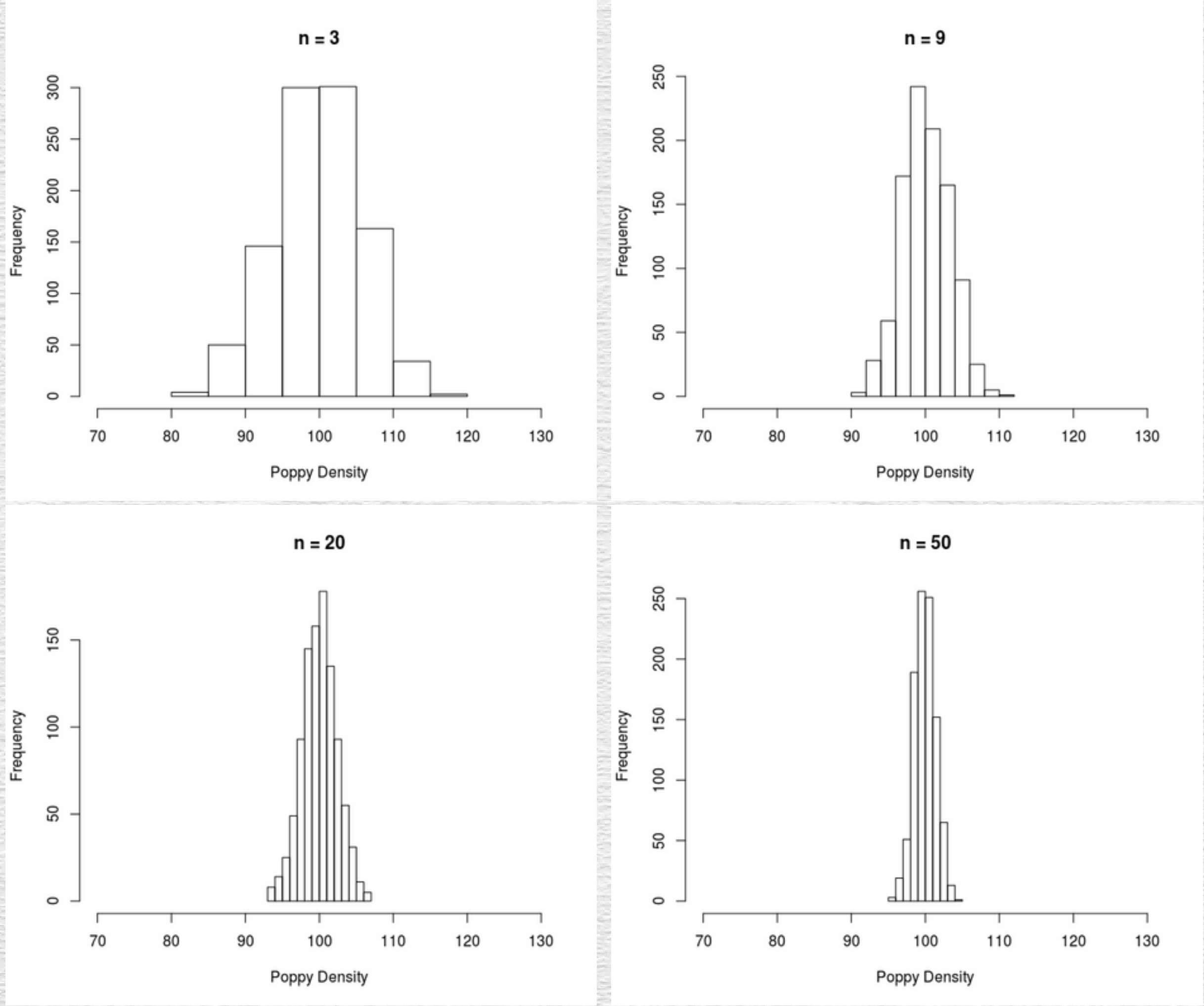
- The *individual* variation among 1 m^2 plots causes sample *estimates* to vary as well
  - Two different sets of 9 plots would have a different set of counts of poppies
  - The counts would therefore have a different mean
- To know how good a single estimate like ours is, we need to understand how estimates tend to vary due to random sampling

https://www.zoology.ubc.ca/~whitlock/Kingfisher/SamplingNormal.htm

Distribution of individual plots



Distribution of sample means = sampling distribution



n = sample size, number of plots counted to obtain the estimated mean

Some generalizations...

- Estimates of means are less variable than individual data values
  - A mean is a measure of **central tendency**, which is the location of the middle of the sample of data
  - Across multiple samples, middles are less variable than individual data values
- Bigger sample sizes lead to less sampling variation
  - Any single estimate is less likely to be far away from the parameter
  - Estimates from repeated samples will thus be closer together
  - More repeatability = better precision

We can estimate variability among **means** from a single sample of data

- A single sample only gives us a single mean
- We want to know how variable many different means sampled from the same population would be
- $s_{\bar{x}}$ = the standard error, measures variability among sample means:

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

- The smaller $s_{\bar{x}}$ is a measure of precision – small $s_{\bar{x}}$ means good precision

Standard error from our nine plots:

- The data:

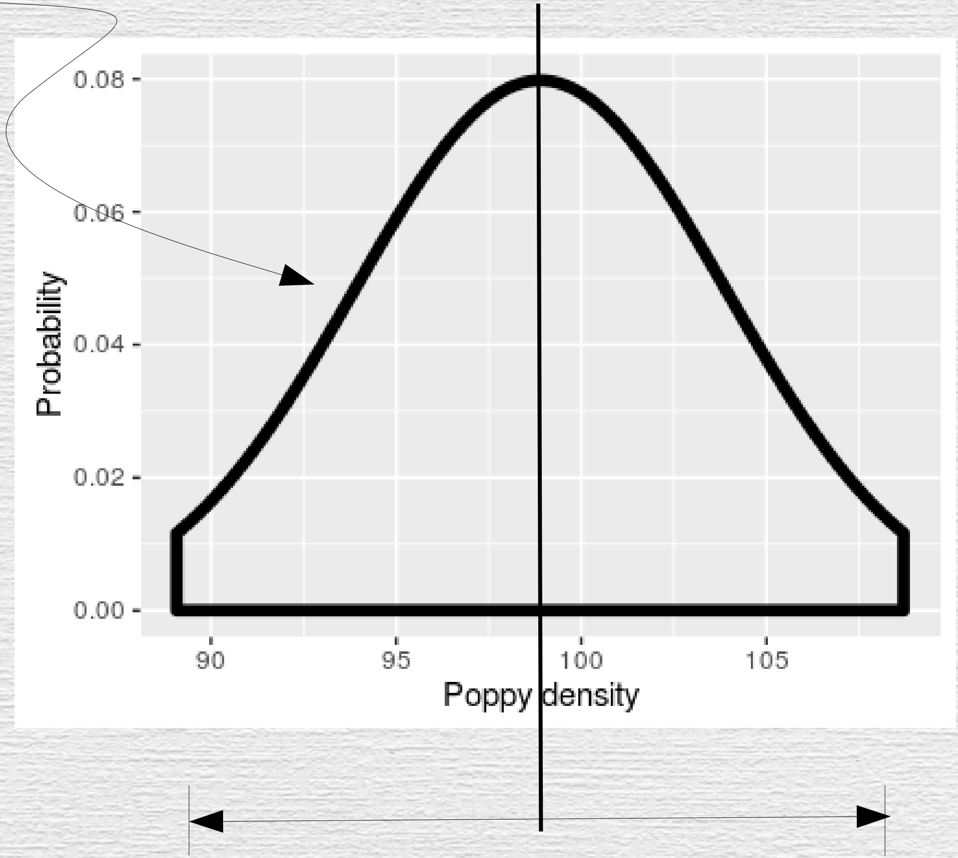
  95, 68, 107, 93, 101, 113, 107, 100, 106
- The average of this sample ($\bar{x}$) is 98.89
- The standard deviation (s) is 13.18
- The sample size (n) is 9
- So, $s_{\bar{x}}$ is 13.18/sqrt(9) = 4.39

Confidence intervals

- Because of random sampling variation, we know:
  - Our estimate of 98.89 poppies/m^2 is probably different from the actual density (μ) by some amount
  - Another sample of 9 will give us a different mean
- We can't know μ for sure, but we can use what we know about random sampling to come up with a range of poppy densities that are good possible values for μ
- We call this range of possible values a **confidence interval**

The basic idea...

The sampling distribution for $\bar{x}$, modeled with t



Lower limit

$\bar{x}$ **= 98.89**

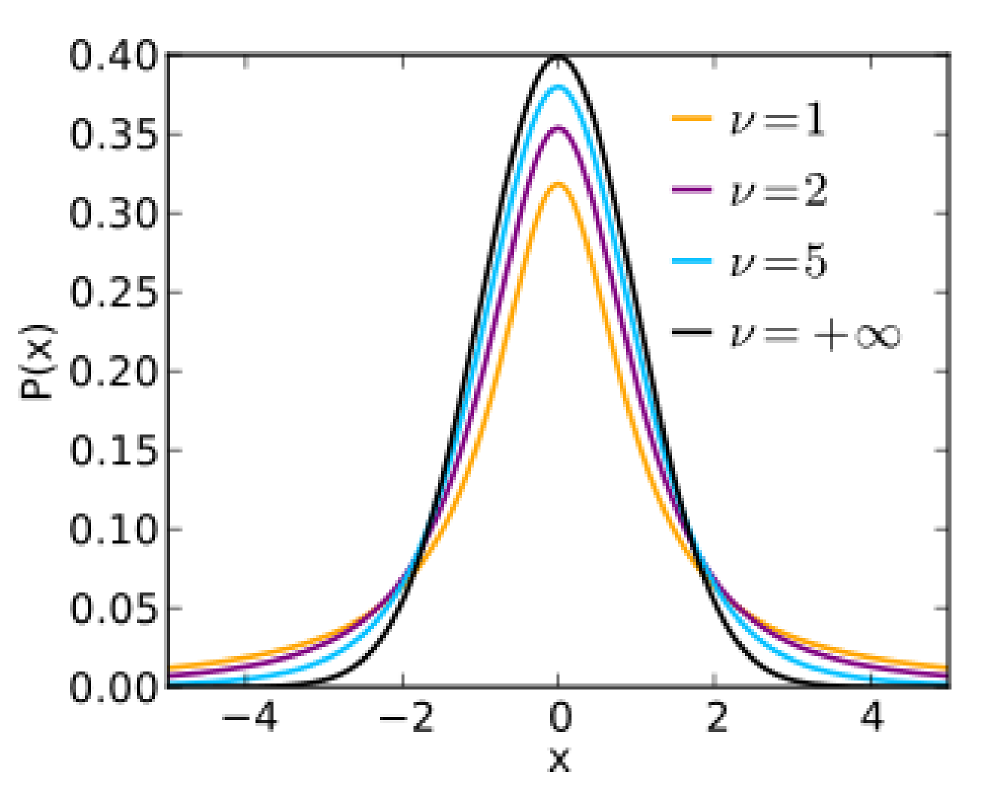
Upper limit

The confidence interval: $\bar{x} \pm ts_{\bar{x}}$

where $ts_{\bar{x}}$ *is called the* ***uncertainty***

The t-distribution

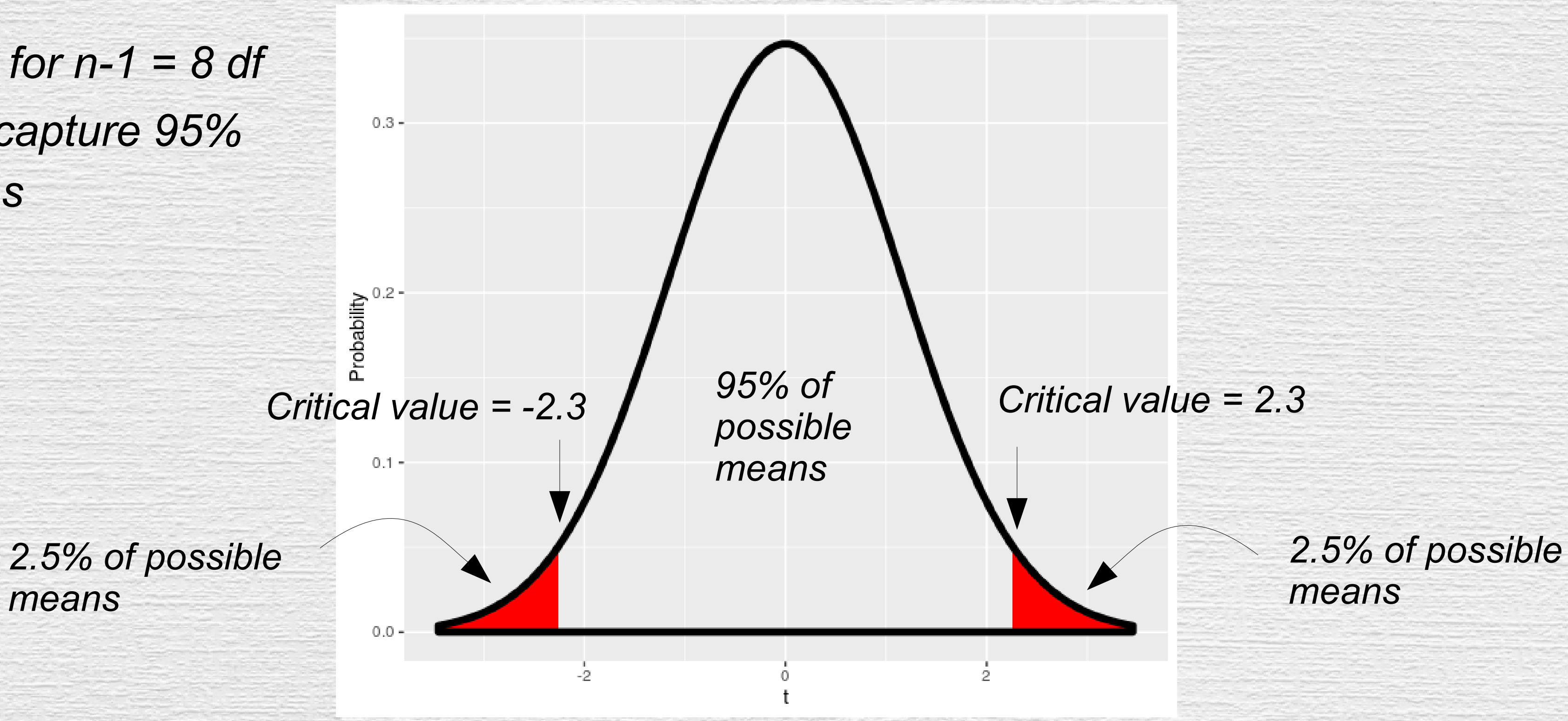
- Similar in shape to the normal, but a better model of random sampling
- The shape depends on degrees of freedom (related to sample size)
- The x-axis is in standard error units



t is used to determine how many $s_{\bar{x}}$ around $\bar{x}$ are needed to include 95% of possible means

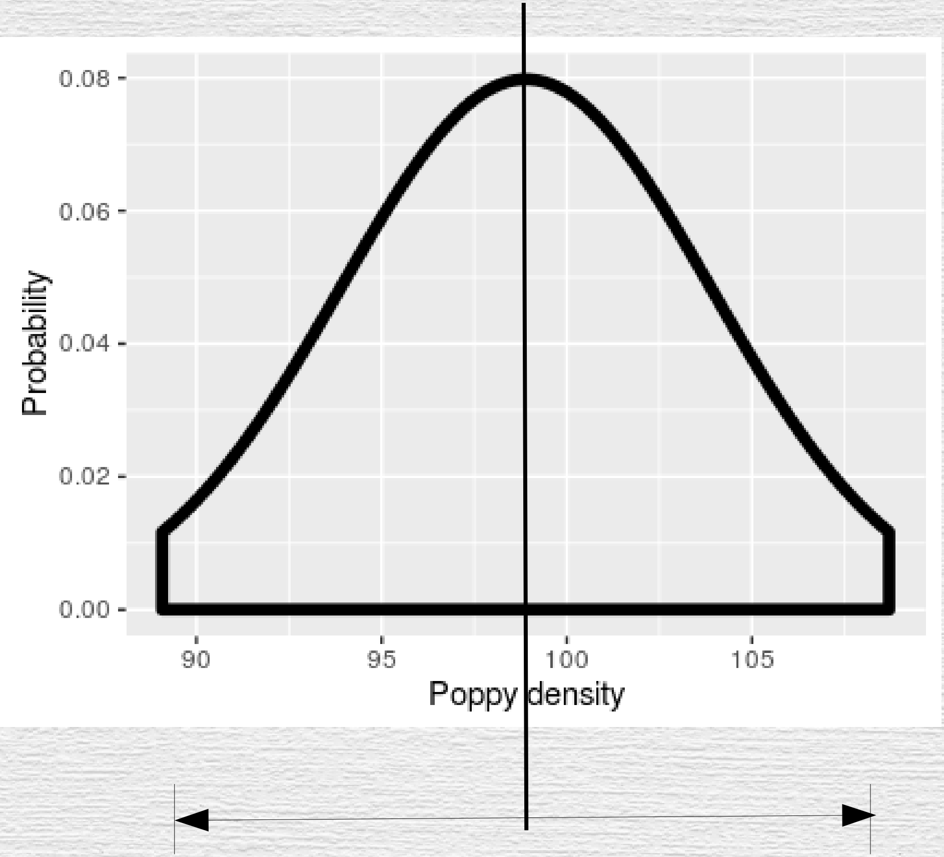
The t-distribution for n-1 = 8 df

Need ± 2.3 $s_{\bar{x}}$ to capture 95% of possible means



Calculations

- An interval is defined by upper and lower limits
- 95% confidence intervals are defined by the upper limit of $\bar{x} + ts_{\bar{x}}$, and the lower limit of $\bar{x} - ts_{\bar{x}}$



$\bar{x}$ = 98.89

Lower limit

98.89 – 2.3 (4.39)

= 88.79

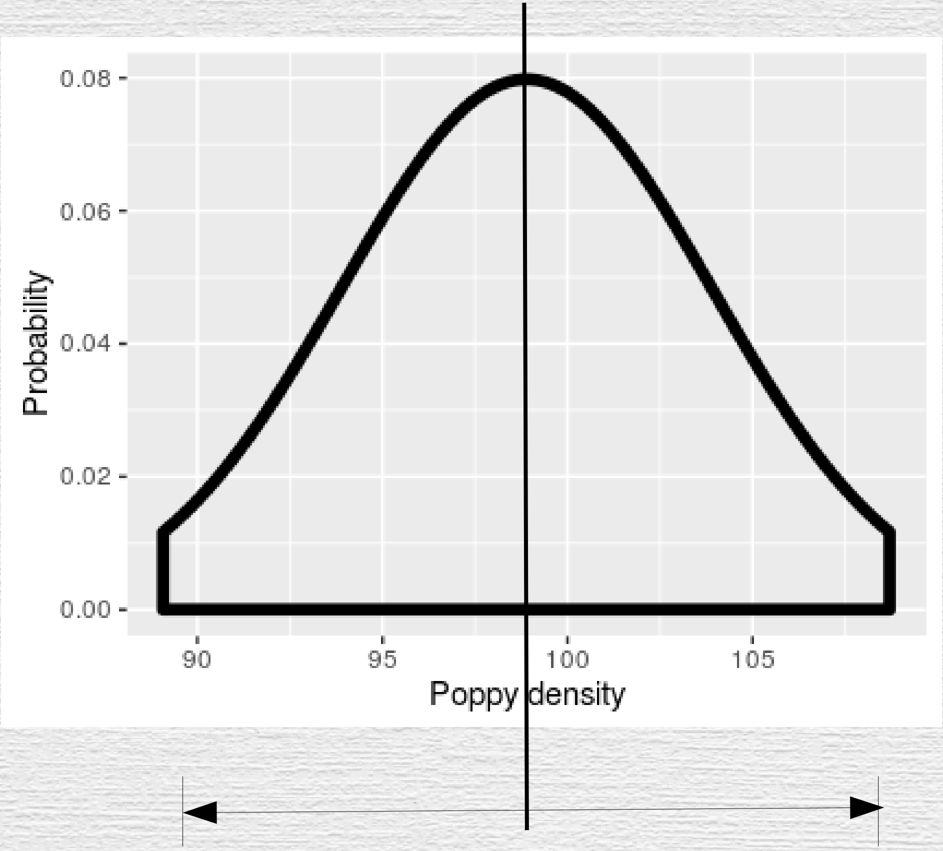
Upper limit

98.89 + 2.3 (4.39)

= 108.99

Interpretation

The estimated poppy density is 98.89 m^{-2}, with 95% confidence that the density is between 88.79 and 108.99



Lower limit = 88.79

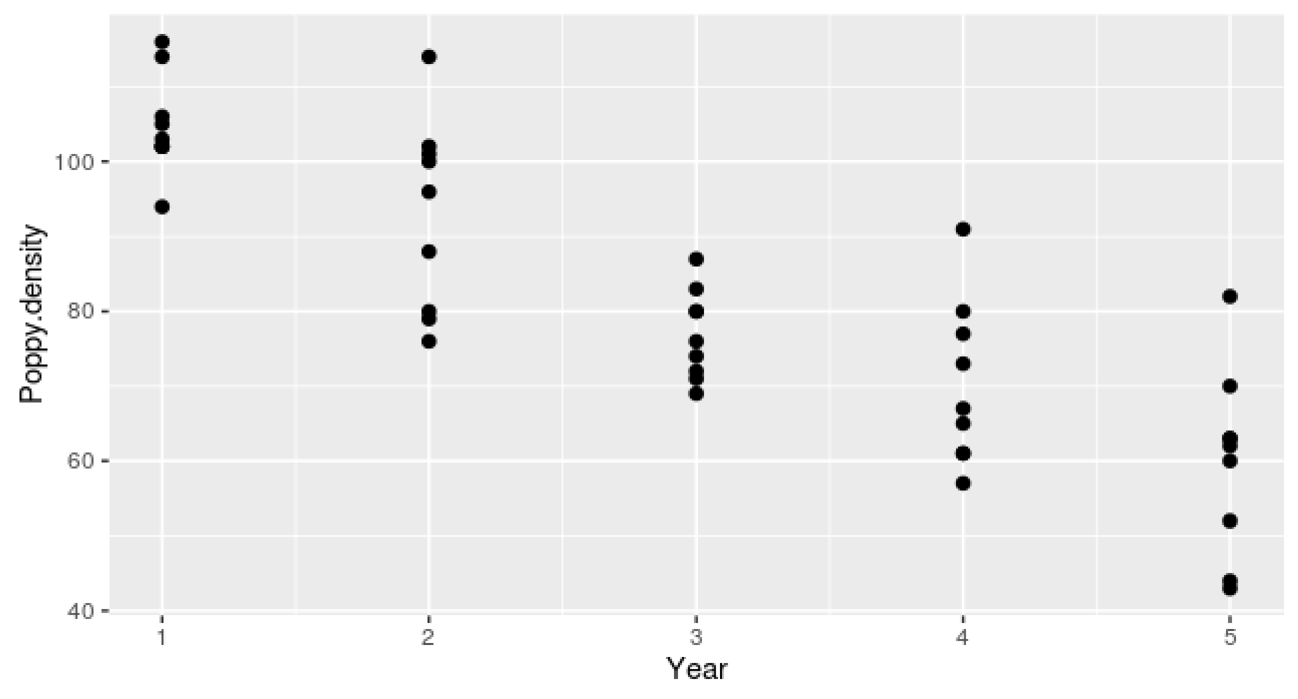
$\bar{x}$ = 98.89

108.99 = Upper limit

Estimation: summary

- We work with samples, but want to generalize about populations
- This is done by estimating population parameters with sample data
- We use the standard error as a raw measure of sampling precision (consistency, repeatability)
- We use the confidence interval to tell us the range of values that are likely to be obtained if we sampled again
  - Since the true population mean is one of the possible sample means, the confidence interval has a 95% chance of including the population mean

So you want to know if poppy density is declining over time...



*Nine **different** plots sampled each year*

What do we want to know?

- Is there a change in number over time?
  - If so, is it a decline or an increase?
- Find the line with equation:

  Poppy density = m (Year) + b

  that fits the data best
  - m = slope = (change in density)/(change in years)
    - Annual rate of change in density
    - If m is negative, then density is (increasing or decreasing?)
  - b = intercept = density at year 0 (i.e. value of y when x = 0)
    - Usually not interpreted – a fitted constant needed for the line to hit the y-axis at the right place so that the line goes through the data
- How do we know what line fits best?

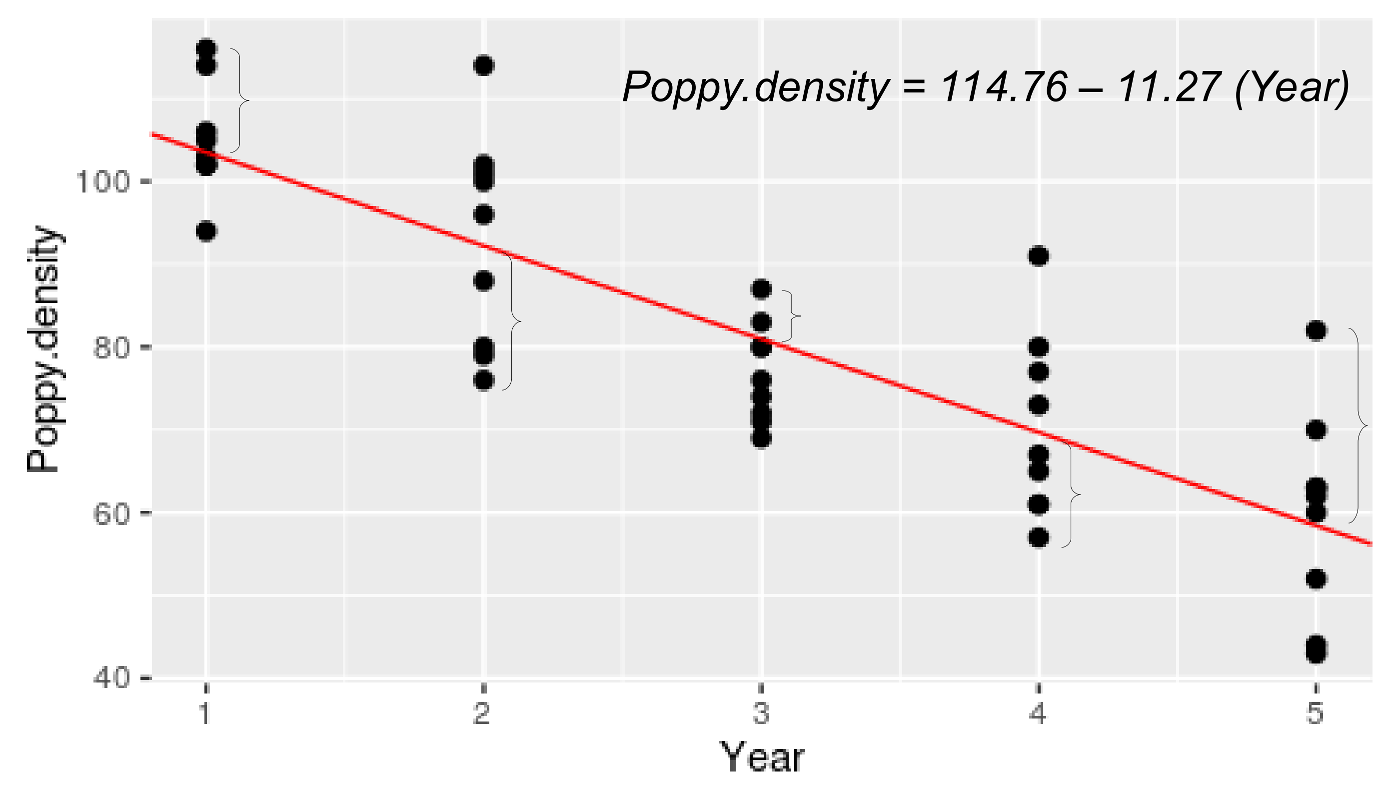
The least squares line

The least squares criterion = the best fit line minimizes the squared residuals

Residuals: data values – predicted values

The slope of -11.27 means that density is decreasing by 11.27 poppies/m^2 each year

LS line is as close as possible to all of the data at once

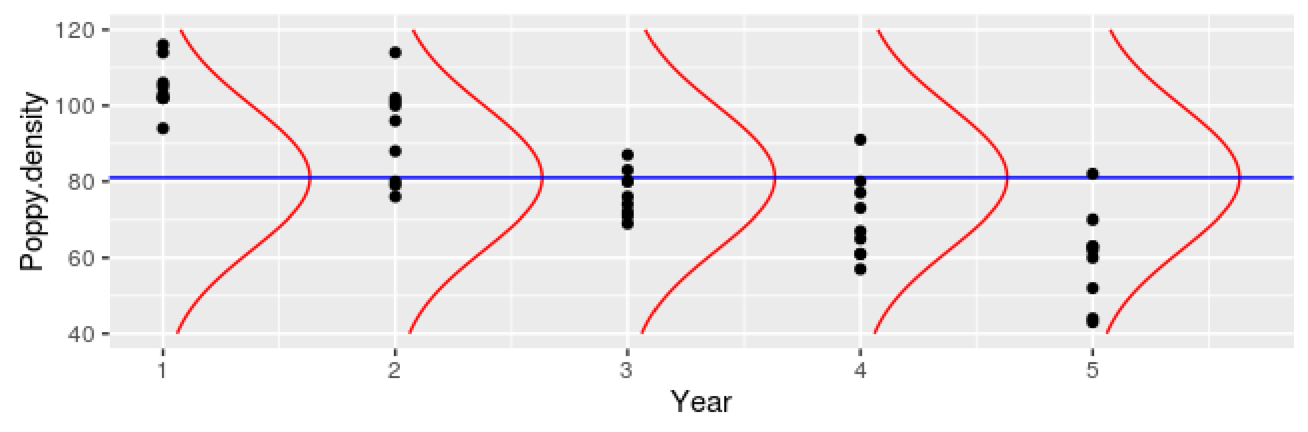


Can we be sure our negative slope represents an actual decline?

- Q: How do we know this apparent decline isn’t just the result of chance?
- A: We can’t be completely certain
  - Randomly generated data can appear to show patterns
- But, we can ask “What is the probability of observing a slope of this size in a sample of data if there actually isn’t a decline in the population?”
- This is a statistical hypothesis, and we evaluate it with a statistical **null hypothesis test**

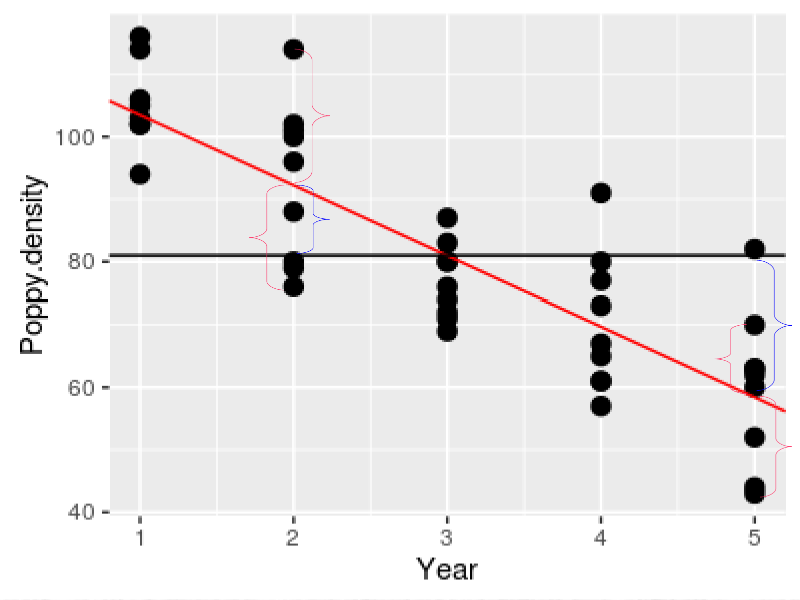
Testing a null hypothesis about the effect of year on poppy density

- Null hypothesis = no effect, no difference, randomness
- No relationship between poppy density and year is a flat line → slope = 0
- Use the probability of getting a line with a slope of -11.267 by sampling a population with a slope of 0 in a test of the null



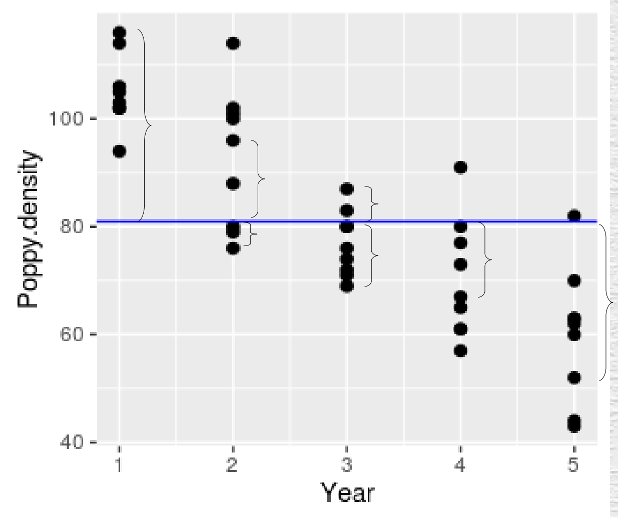
Partitioning the variance

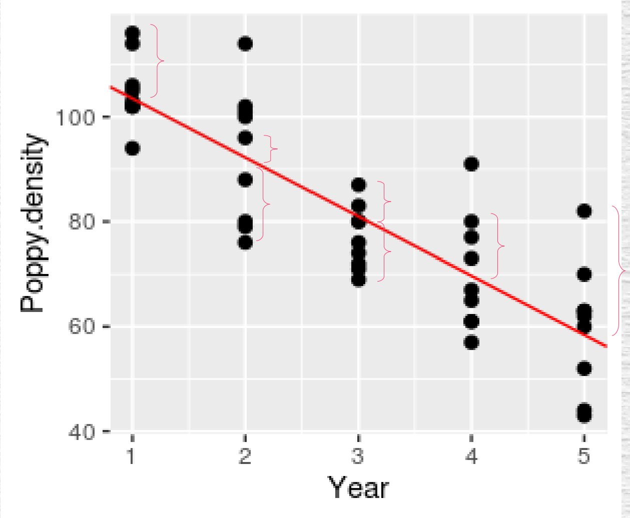
- Think of each data point as being due to the sum of two things:
  - The average poppy density in a given year (blue brackets)
  - Random individual variation around the annual average (red brackets)
- Variation around the horizontal line = total variation
- Divide this into two components:
  - Explained variation = the regression line
  - Unexplained variation = the residuals

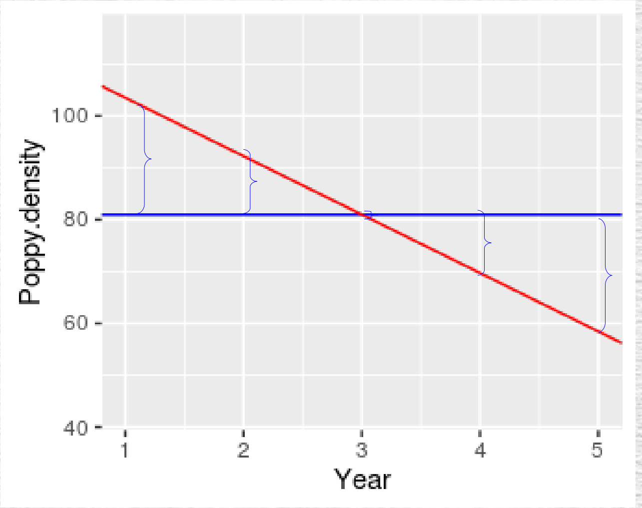


- If the variation explained by the line is large compared to the random variation, then we have reason to think there is a real decline

SST = SSE + SSR







Total SS

Sum of squared differences between Y data and mean of Y data

Residual SS

Sum of squared residuals around the line

Random, unexplained variation

Regression SS

Sum of squared differences between mean of Y and predicted value

Explained variation

Convert SS to variances

- We want to know how explained variation compares with unexplained variation, but SS are totals
  - Each individual data point contributes to the error SS
  - The line is defined by two paraters, which is the basis for the model SS
- Need to convert raw SS into values that can be compared
- Variance is an average amount of variation per degrees of freedom:
- We can convert each of our SS to variances if we divide them by an appropriate degrees of freedom

$$s^2 = \sum \frac{x_i - \bar{x}}{n-1} = \frac{SS}{df}$$

- If the null is true, then both of these variances are actually estimating random variation → should be about the same size

SS and df for each component

- Total degrees of freedom = n – 1 = 44
- Model degrees of freedom is 1 (slope estimate consumes 1 degrees of freedom)
- Residual degrees of freedom is total – model = 44 - 1 = 43
- Each component's MS is calculated as its SS/df

$MS_{total} = SS_{total}/df_{total}$

$MS_{model} = SS_{model}/df_{model}$

$MS_{residual} = SS_{residual}/df_{residual}$

Using MS to calculate the F test statistic

*We need a **test statistic** that measures what is observed in the data*

F is a ratio of two variances (any two variances)

For a regression, we calculate F as:

$F = MS_{model}/MS_{residual}$

If the null hypothesis is true, both MS estimate random variation, and the ratio should be 1

If the null hypothesis is false, model MS will be bigger than residual MS, and F will be bigger than 1

Assemble into an ANOVA table

```
Response: Poppy.density
          Df  Sum Sq Mean Sq F value    Pr(>F)
Year       1 11424.4 11424.4  113.26 1.262e-13
Residuals 43  4337.5   100.9
```

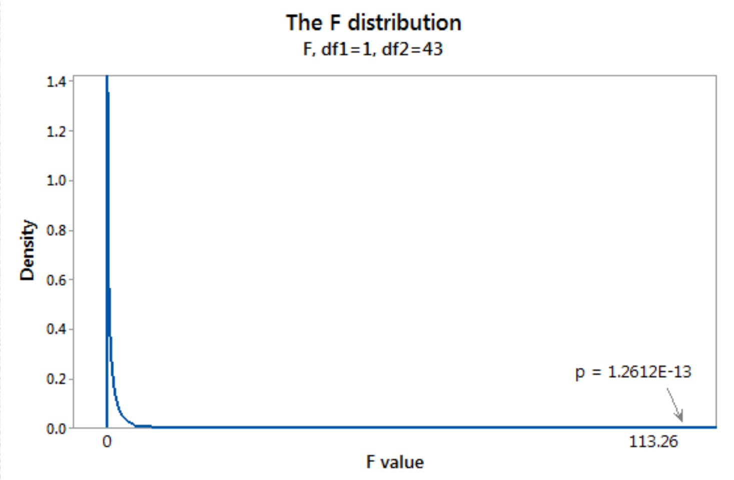
Sampling distribution for F is the F distribution

Shape is determined by both numerator (1) and denominator (43) degrees of freedom

The probability of an F of 113.26 or greater if the null is true is area under curve from 113.26 to ∞

$p = 1.262 \times 10^{-13}$ – very small

Reject the null hypothesis, conclude that the slope is not 0 – the regression is statistically significant



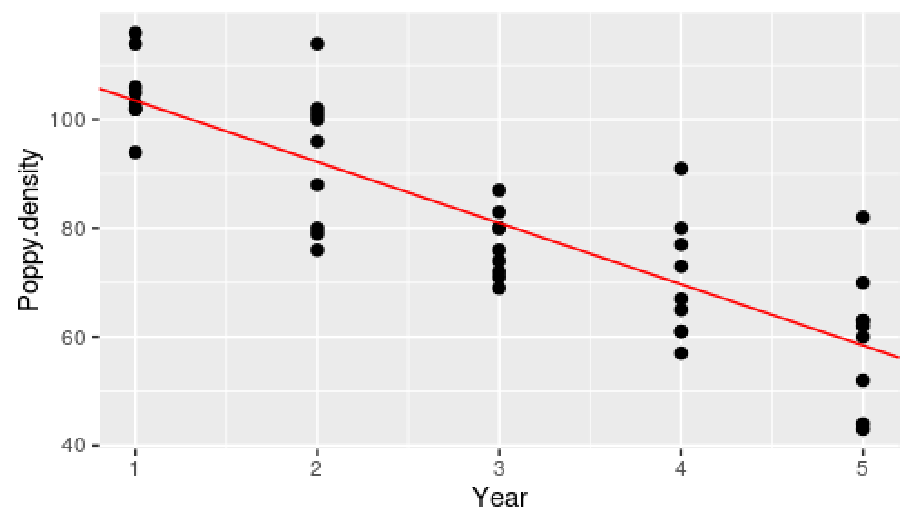
Strength of the relationship - r^2

```
Response: Poppy.density
          Df  Sum Sq Mean Sq F value    Pr(>F)
Year       1 11424.4 11424.4  113.26 1.262e-13
Residuals 43  4337.5   100.9
```

r^2 = coefficient of determination

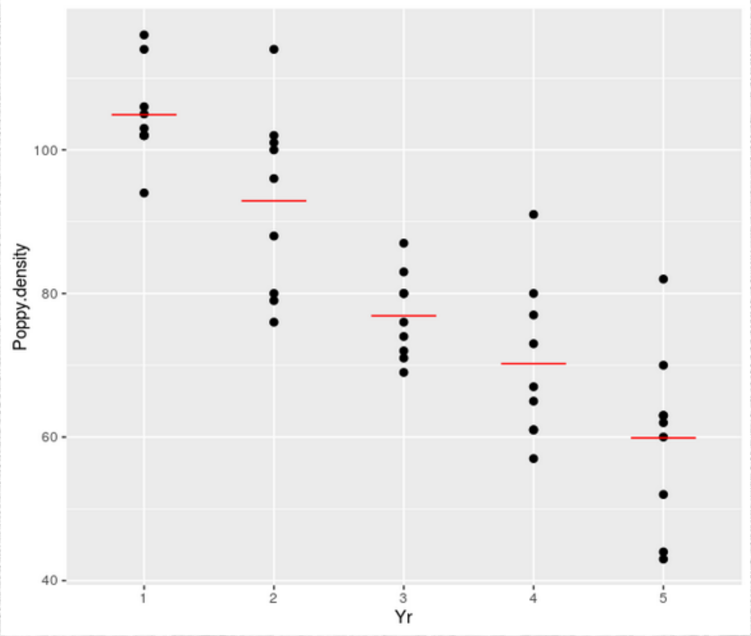
Measures the proportion of total variation explained by the line

$$r^2 = \frac{SS_{\text{regression}}}{SS_{\text{total}}} = \frac{11424.4}{11424.4 + 4337.5} = 0.72$$

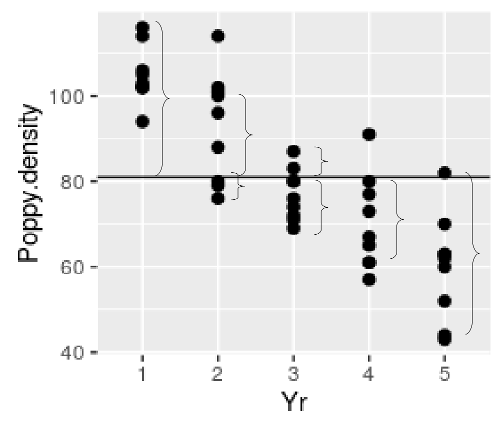


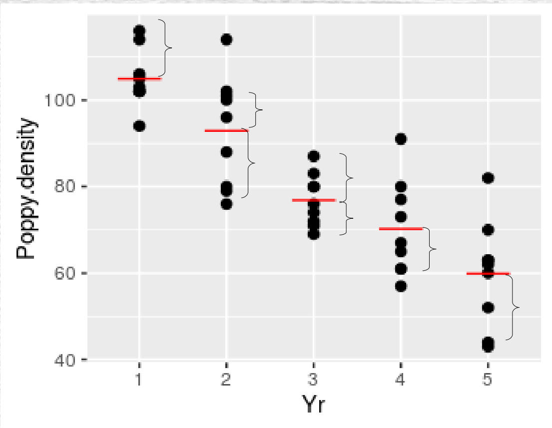
We could instead treat year as a category

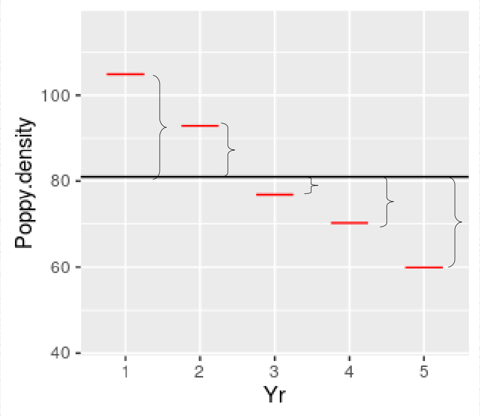
- We could treat these as grouped data, with year representing the groups
- We would then ask if the means are different from one another
- The null hypothesis would be that all the group means are the same



SST = SSE + SSF







Total SS

Sum of squared residuals, using the mean of the Y data

Error SS

Sum of squared residuals, using the group means

Random, unexplained variation

Factor SS

Sum of squared differences between mean of Y and group means

Explained variation

The ANOVA table

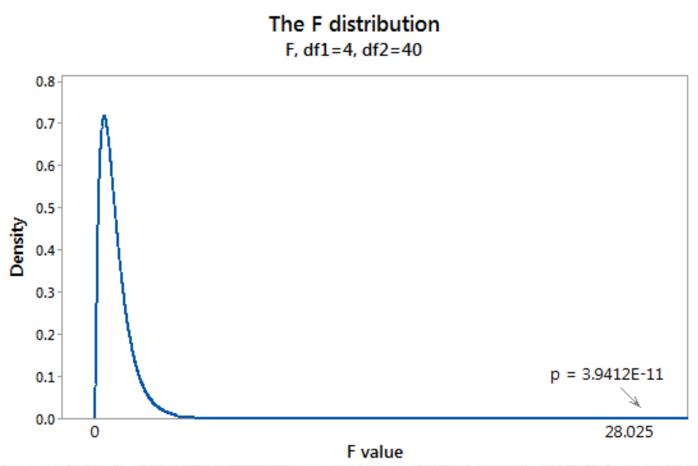
```
Response: Poppy.density

          Df  Sum Sq Mean Sq F value    Pr(>F)
Yr         4 11616.8 2904.20  28.025 3.941e-11
Residuals 40  4145.1  103.63
```

DF:

Predictor (Yr) gets number of groups – 1 = 4

Total is sample size – 1 = 44

Residual is $DF_{total} – DF_{yr}$ = 44 – 4 = 40

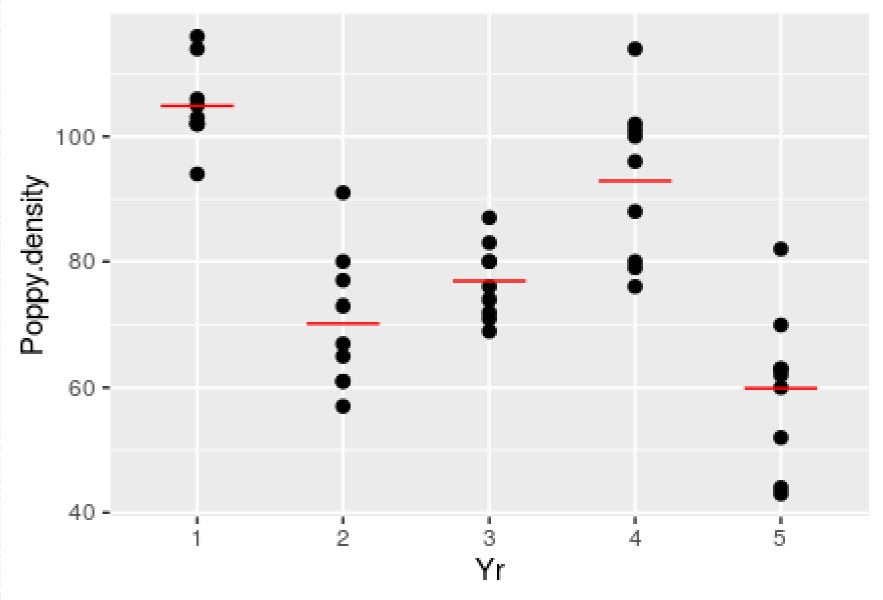


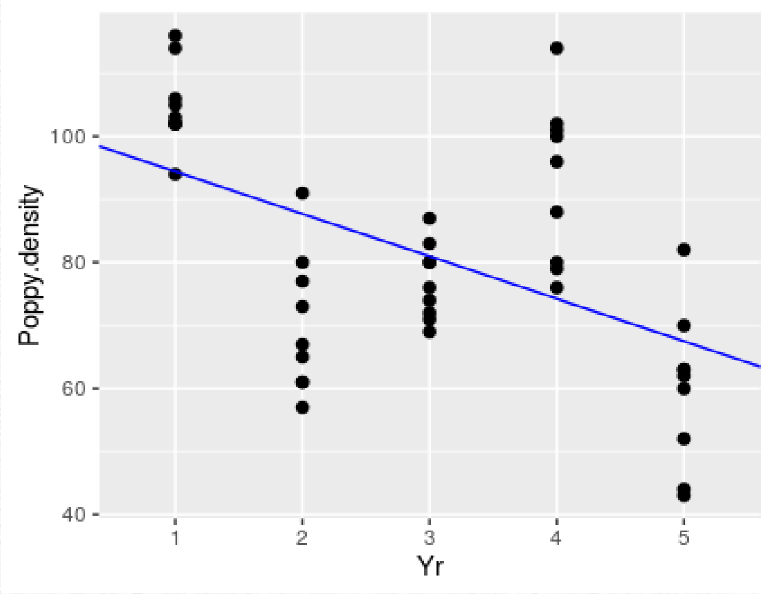
The p-value tells us that the probability of this amount of difference between means if all years are the same is 1.262 x 10^{-13}

Which to use?

- Clearly they are very similar
- The analysis is more powerful (i.e. more likely to detect a real change) with more residual DF
  - Regression has the advantage – only 1 df for Year, which left 43 residual
  - ANOVA needed 4 for Yr, left only 40 residual
- But, regression is only a better choice if the decline is linear

What about these data?



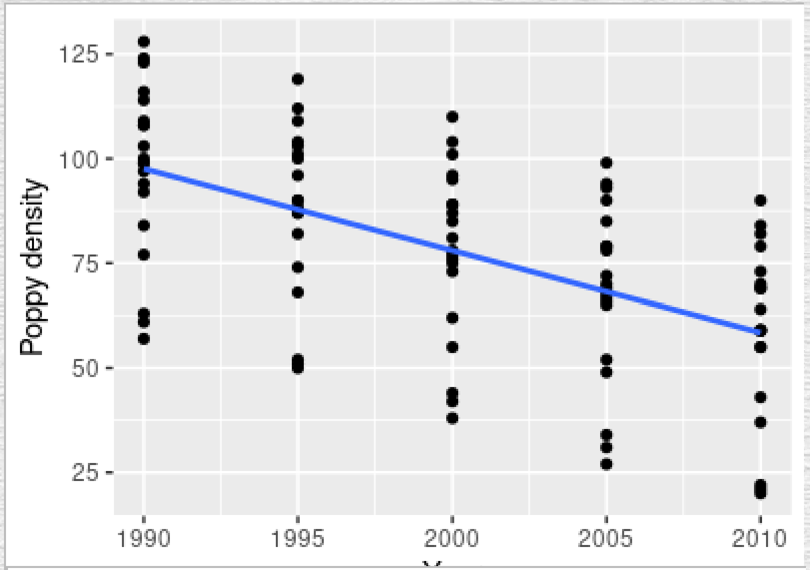


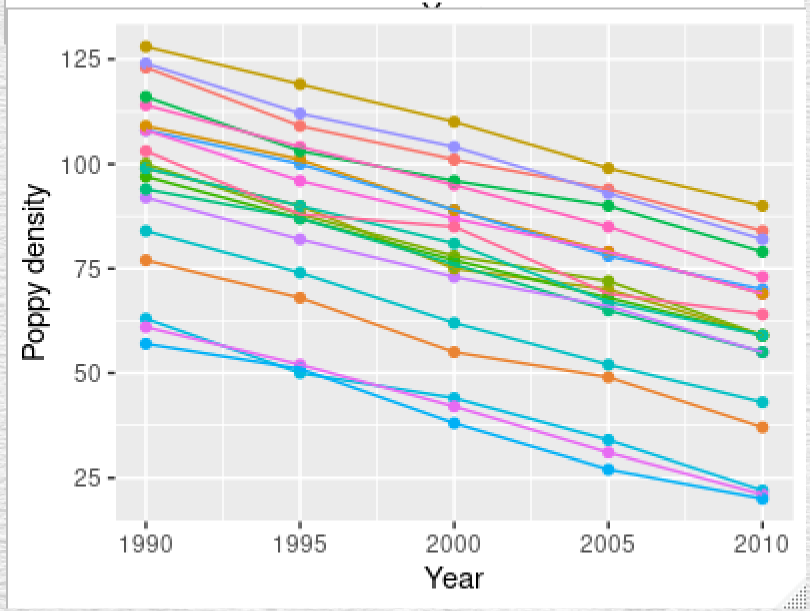
Regression and ANOVA summary

- We use regression and ANOVA to analyze how a predictor affects a numeric response (poppy density)
  - In regression the predictor is also numeric (year treated as a number)
  - In ANOVA, the predictor is a category (year treated as a grouping variable)
- We test a null hypothesis about the chances of our results occurring at random
- Regression is a better choice than ANOVA, when the response is linear

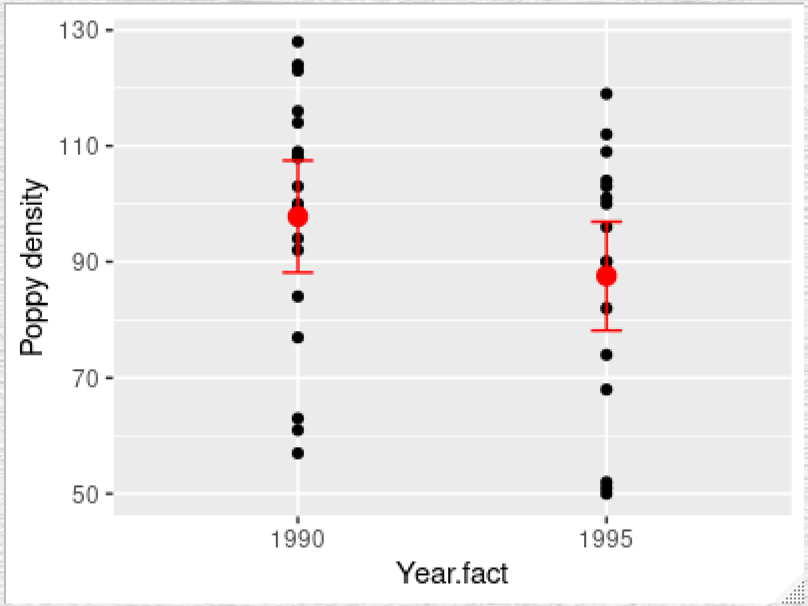
Repeated measures data

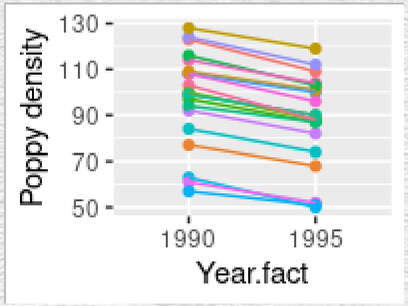
- It is often beneficial to record conditions at the same locations every year
- More sensitive to change, because the differences at individual points is the focus
- But, to get the benefit of the design must also analyze the data as paired data

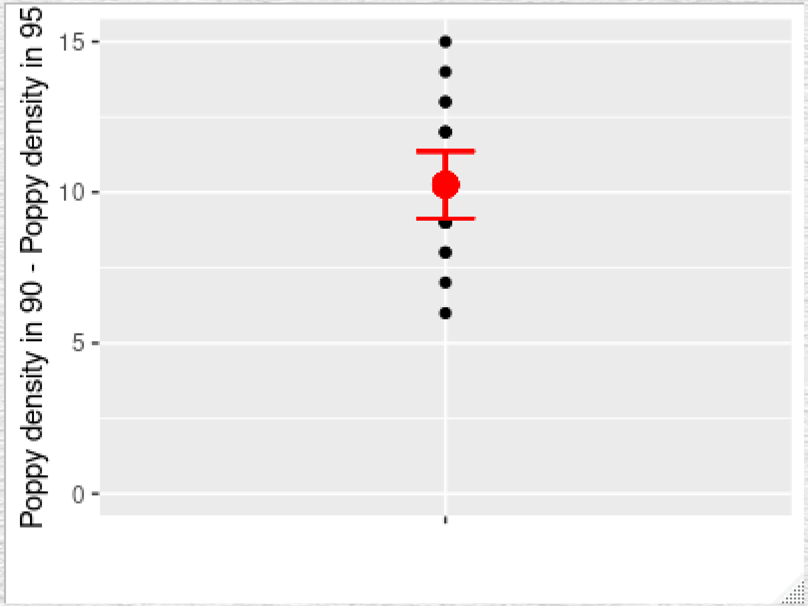




Simplified example: two time points







If the pairing isn't accounted for, difference between means isn't statistically significant (p = 0.12)

Using differences between measures of same points at two times gives very consistent changes – signficantly different from 0 (p = 6.3x10^{-14}*)*