Outline for Today

1. Review of In-class Exercise
2. Bivariate hypothesis testing 2: difference of means
3. Bivariate hypothesis testing 3: correlation

Task for Next Week

- Any questions?

In-class Exercise

```
 Is regime | GDP per capita (US$):
         a |       2 cats
democracy? |       Low       High |     Total
-----------+----------------------+----------
        No |        49         19 |        68
       Yes |        40         69 |       109
-----------+----------------------+----------
     Total |        89         88 |       177

          Pearson chi2(1) =  20.9459   Pr = 0.000
```

Interpreting Stata Output

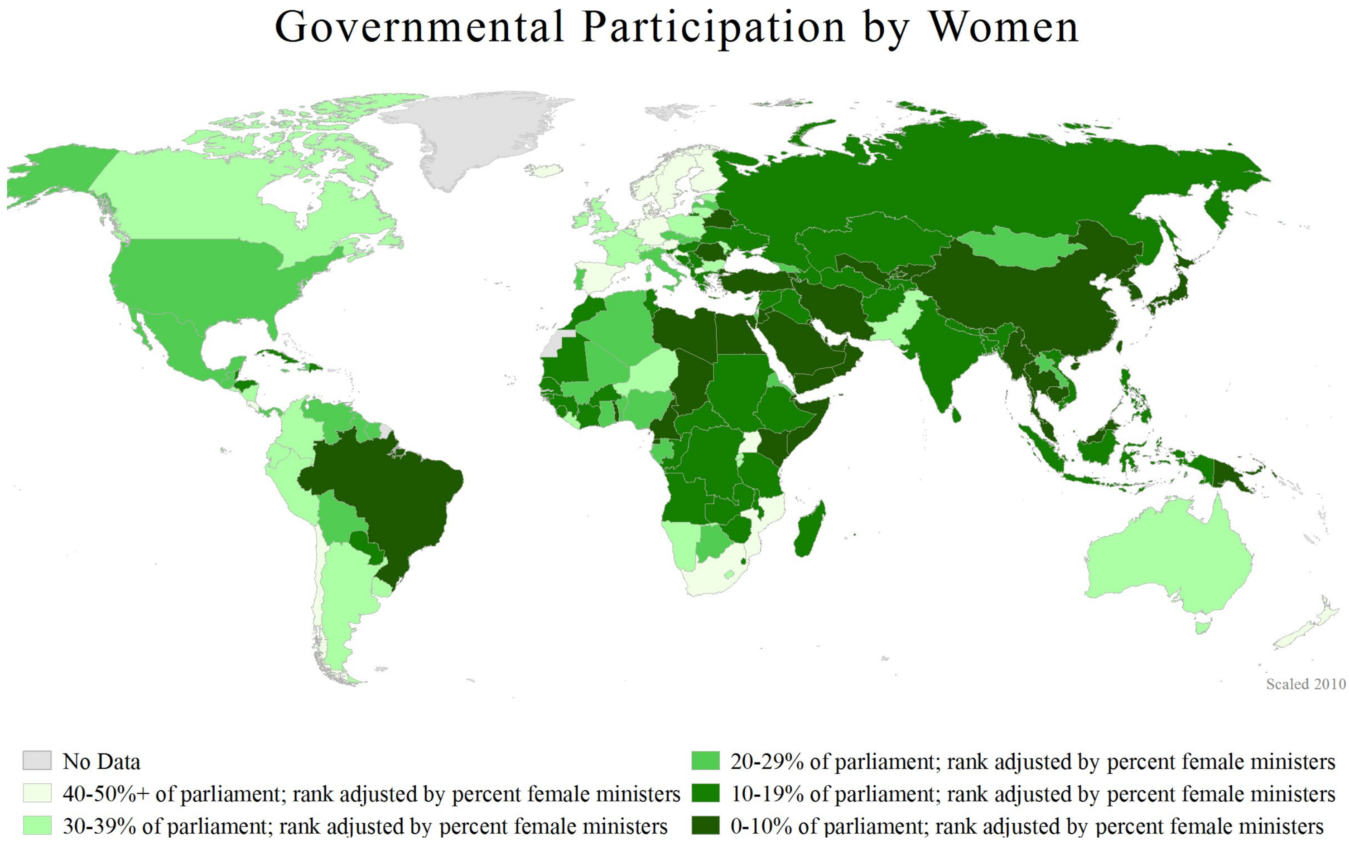
- p-value will always be between 0 and 1
- When Stata gets a p-value smaller than 0.0005, it will round the number and tell you it's 0.000.
- When Stata tells you it's 0.000, what it actually means is $0 < p < 0.0005$.
- When Stata tells you it's 0.000, we say p-value is smaller than 0.001.

Bivariate Hypothesis Test 2: Difference of Means

- Y is continuous but X is categorical

- We follow the same logic and same steps as cross-tabulation analysis:
  1. Form the null and alternative
  2. Examine and describe the sample
  3. Compare the observed and expected
  4. Reject or not reject the null

Female Representation in Parliament



Female Representation and PR System

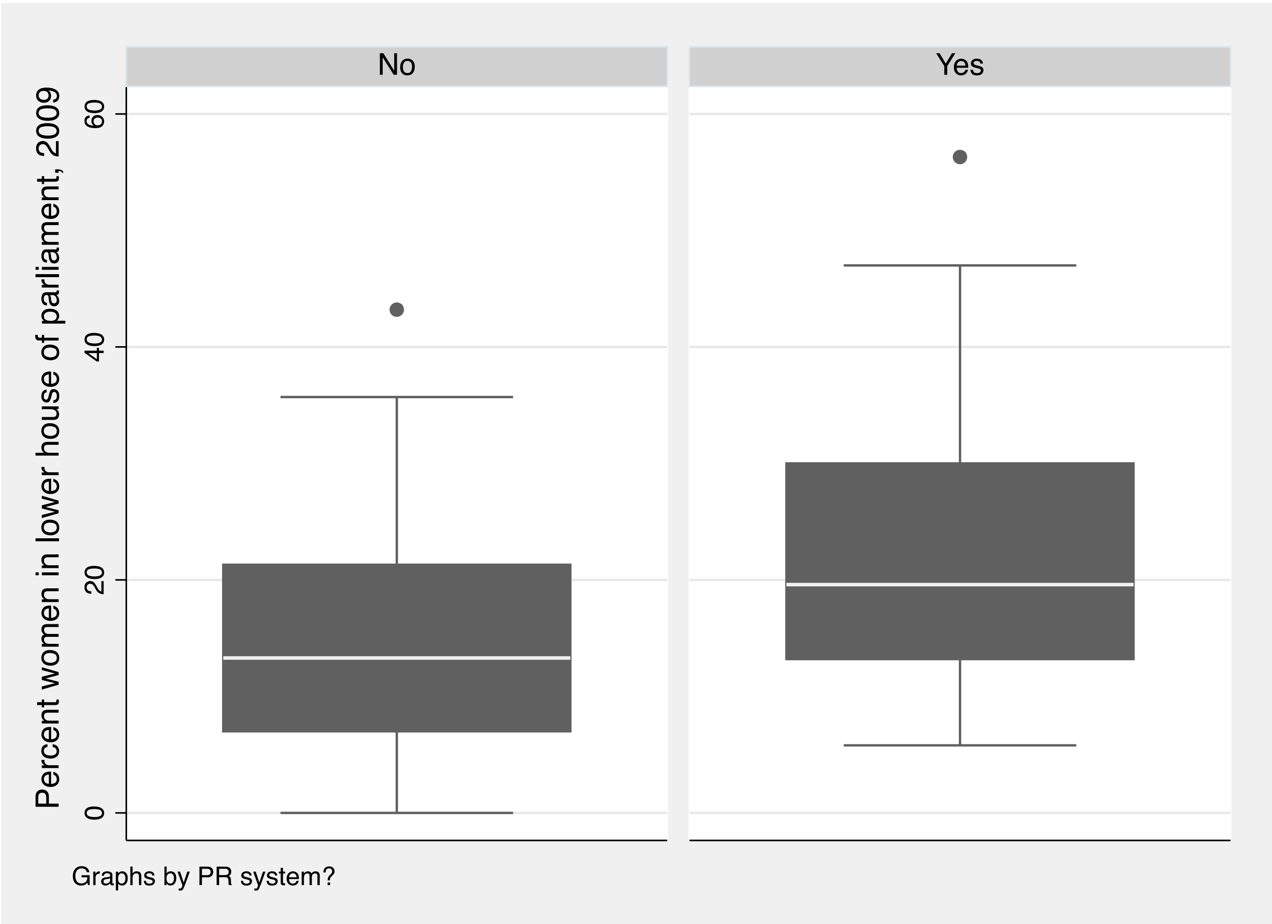
Proportional representation voting system, compared with majority voting system, favours minority and under-represented groups in a society.

Hypothesis

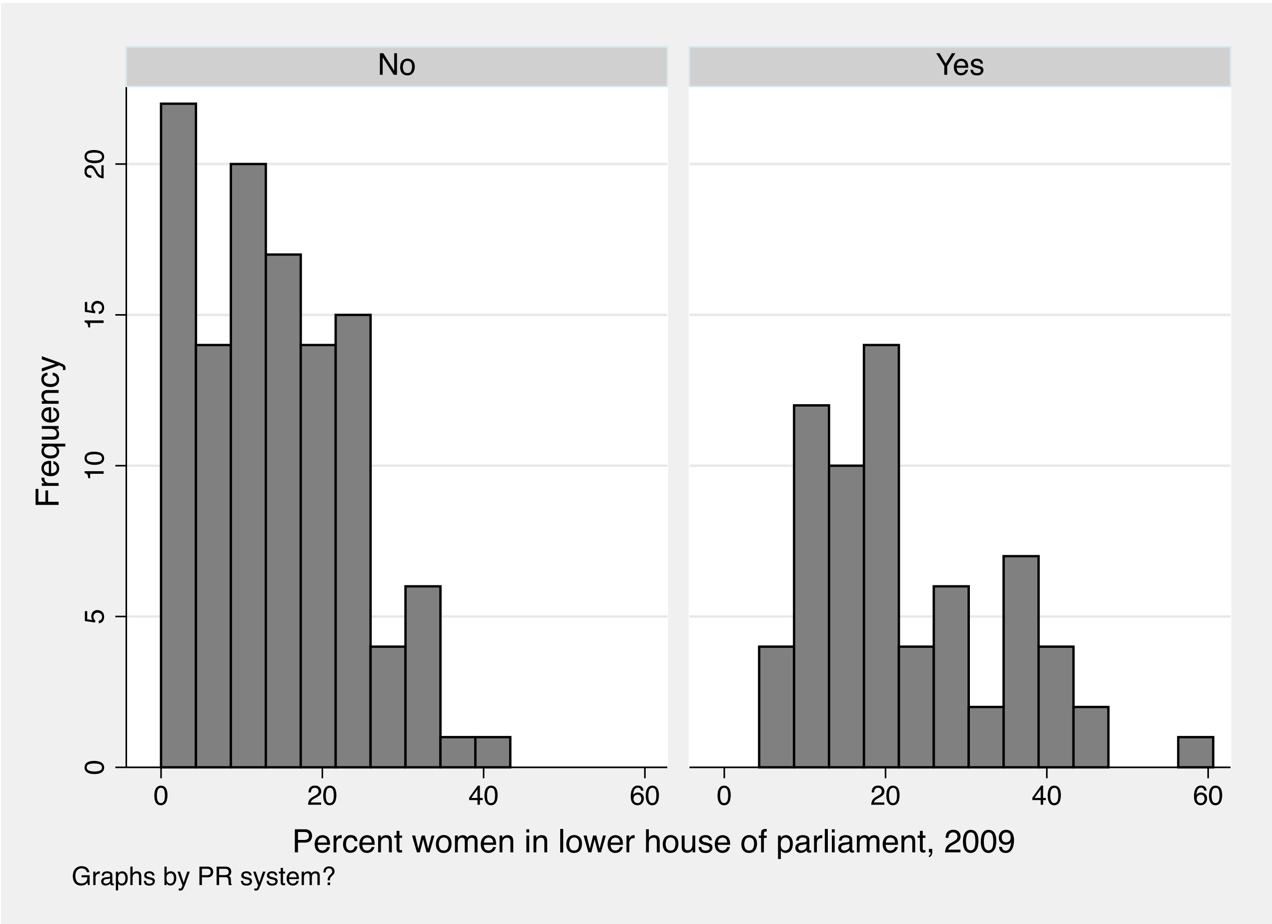
Female representation is higher in countries that adopt the PR system than in countries that adopt the majority system.

- What's Y ? Female representation (% women in parliament)
- What's X ? Whether a country has a PR system or not (Yes or No)
- The null hypothesis: there is no relationship between PR system and female representation

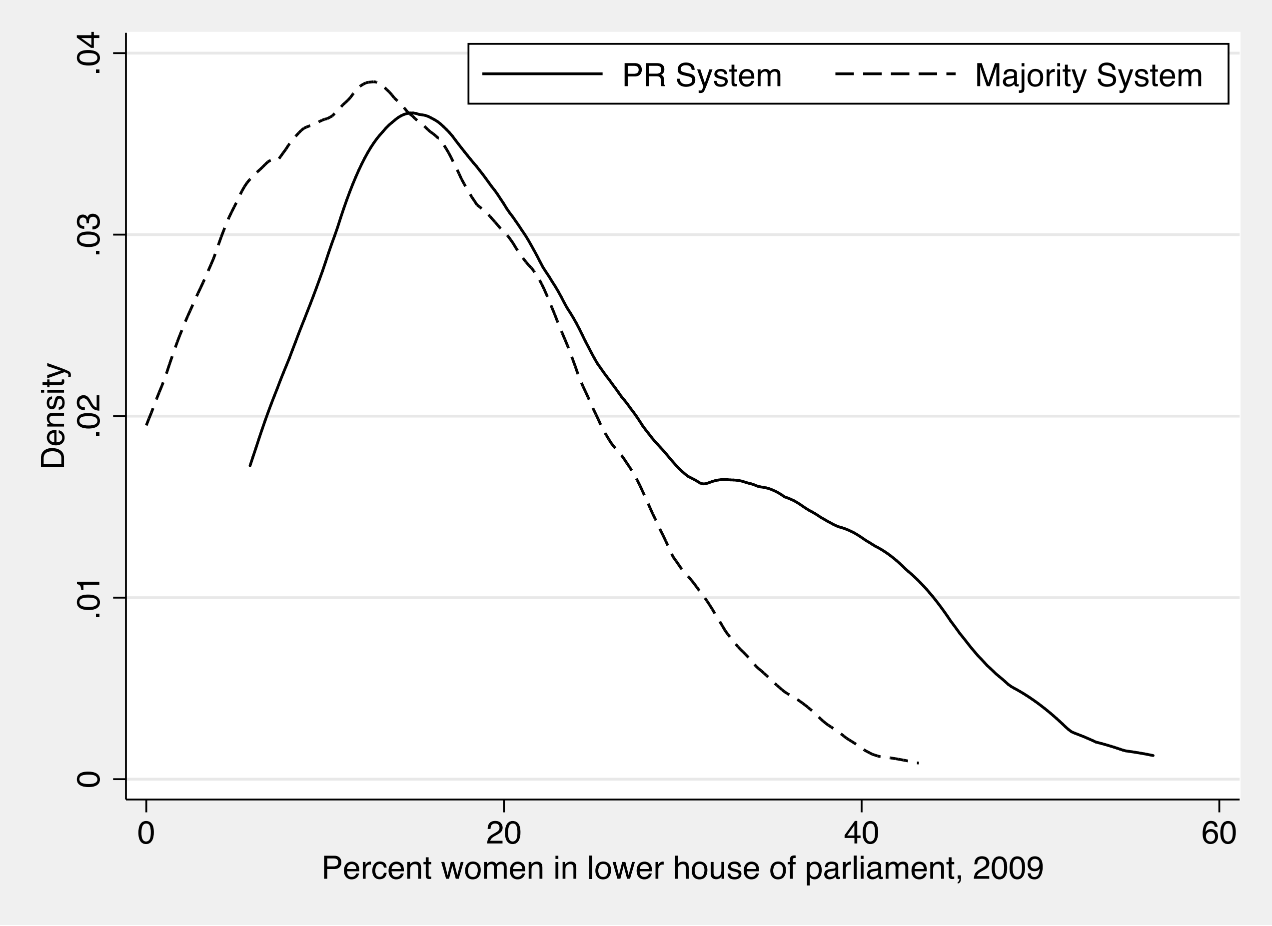
Graphical Summaries of Y by X



Graphical Summaries of Y by X



Graphical Summaries of Y by X



Numerical Summaries of Y by X

```
bysort pr_sys: sum women09
-> pr_sys = No

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     women09 |       114    14.15965    9.459815          0       43.2

-----------------------------------------------------------------------
-> pr_sys = Yes

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     women09 |        66    22.38939    11.71783        5.8       56.3

-----------------------------------------------------------------------
. sum women09

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     women09 |       180    17.17722    11.05299          0       56.3
```

Calculating the Test Statistic

- It does seem that there is some relationship between X and Y in the sample.

- The next step is to see if this observed relationship is sufficiently different from the relationship we would obtain if the null hypothesis were true.

- This involves calculating the test statistic:
  - In univariate analysis of the mean, we calculated the sample mean.
  - In cross-tabulation, we calculated the χ^2 statistic.
  - In difference of means test, we calculate the t-statistic.

t-Statistic

The t-statistic for the difference of means test is

$$t = \frac{\bar{Y}_1 - \bar{Y}_2}{se(\bar{Y}_1 - \bar{Y}_2)},$$

where

- $\bar{Y}_i$ is the sample mean of Y for group i
- In our example, $\bar{Y}_{PR}$ is the mean percentage of women in parliament for countries with PR system, and $\bar{Y}_M$ is the mean percentage of women in parliament for countries with Majority system.
- t is small (in absolute values) when the difference between two mean values are similar.
- The greater the se, the smaller the t-statistic $\rightarrow$ less confidence we have in rejecting the null.

Standard Error of the Difference

Recall that, in the univariate case, *se* of the sample mean is:

$$se(\bar{Y}) = \frac{s}{\sqrt{n}}.$$

In the bivariate case, *se* of the difference of sample means is:

$$se(\bar{Y}_1 - \bar{Y}_2) = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}},$$

where

- n_i is the sample size for group i
- s_i is the standard deviation for group i

Standard Error of the Difference

```
 -> pr_sys = No

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     women09 |       114    14.15965    9.459815          0       43.2

-----------------------------------------------------------------------
-> pr_sys = Yes

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     women09 |        66    22.38939    11.71783        5.8       56.3
```

Let's say group 1 is Majority System and group 2 is PR System:

- $n_1 = 114,\ \bar{Y}_1 = 14.15965,\ s_1 = 9.459815$
- $n_2 = 66,\ \bar{Y}_2 = 22.38939,\ s_2 = 11.71783$
- $se(\bar{Y}_1 - \bar{Y}_2) = \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}} \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$

$$= \sqrt{\frac{(114-1)9.459815^2+(66-1)11.71783^2}{114+66-2}} \times \sqrt{\frac{1}{114} + \frac{1}{66}} = 1.5995682$$

t-Statistic

- $n_1 = 114$, $\bar{Y_1} = 14.15965$, $s_1 = 9.459815$
- $n_2 = 66$, $\bar{Y_2} = 22.38939$, $s_2 = 11.71783$
- $se(\bar{Y_1} - \bar{Y_2}) = 1.5995682$

$$t = \frac{\bar{Y_1} - \bar{Y_2}}{se(\bar{Y_1} - \bar{Y_2})} = \frac{14.15965 - 22.38939}{1.5995682} = -5.144976$$

We now need to determine how unusual (significant) the t-statistic of -5.145 is.

Signs of t-Statistic

As we calculated the t-statistic by setting Majority System as group 1 and PR system as group 2,

- Our causal theory expects $Y_1 - Y_2$ to be negative, and hence t-statistic < 0.
- The null hypothesis expects $Y_1 - Y_2$ to be zero, hence t-statistic to be 0.
- The smaller the t statistic, the more confidence we have in our causal theory.

Stata does not know which alternative hypothesis you have:

- $\bar{Y}_1 > \bar{Y}_2$
- $\bar{Y}_1 \neq \bar{Y}_2$
- $\bar{Y}_1 < \bar{Y}_2$

so it reports all three results.

Sampling Distribution

Recall that

- the sampling distribution of sample mean follows Normal distribution;
- the sampling distribution of χ^2 statistic follows χ^2 distribution.

Similarly, the sampling distribution of t-statistic follows (Student's) t-distribution.

Let's learn a little bit about probability distributions.

Probability Distribution

Probability distribution: list of probabilities assigned to possible outcomes.

One way to describe a probability distribution is to identify PMF or PDF:

- Discrete ($\simeq$ categorical) variables: probability mass function (PMF)
  - Bernoulli distribution: e.g., heads with p, tails with $1 - p$
- Continuous variables: probability density function (PDF)
  - uniform distribution
  - Normal distribution
  - χ^2 distribution
  - t distribution

  Area under the curve represents the probabilities.

Coin Flips: Bernoulli Distribution

Let's say the outcome of a coin flip is our variable of interest:

- Outcomes: Heads or Tails
- Probabilities: Heads with p, Tails with $1 - p$
- When we have a variable that has two possible outcomes, the probability distribution is called Bernoulli distribution.

One probability distribution is

$$\begin{cases} \textit{Heads} & 0.5 \\ \textit{Tails} & 0.5 \end{cases} \tag{1}$$

Another probability distribution is

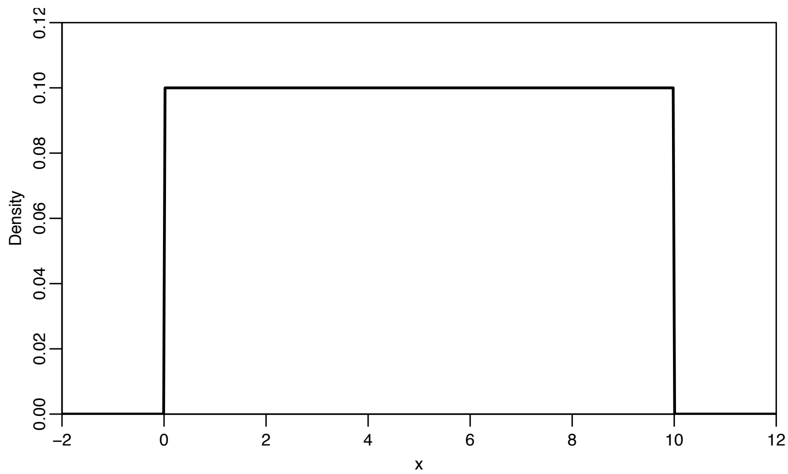
$$\begin{cases} \textit{Heads} & 0.1 \\ \textit{Tails} & 0.9 \end{cases} \tag{2}$$

Both (1) and (2) are proper PMFs, as they list up all possible outcomes as well as the associated probabilities.

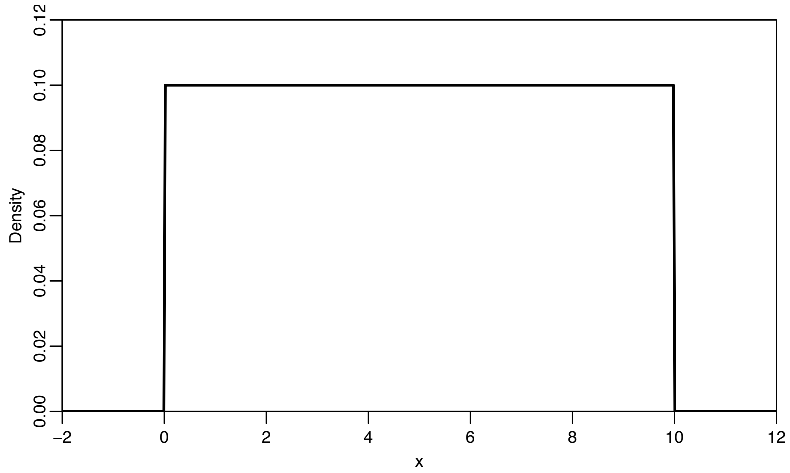
Uniform Distribution

Let's say a random variable x is distributed continuously from 0 to 10:

- Outcomes: any number between 0 and 10.
- As there are infinite number of values between 0 and 10, we cannot list up all values and associated probabilities.
- Instead, we describe the probability distribution with a graph of PDF.



Uniform Distribution



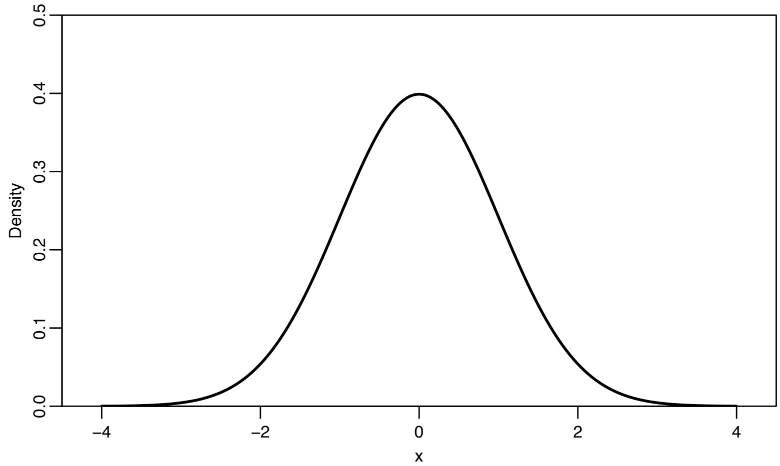
With a graph of PDF, we can calculate the probabilities that x takes a certain range of values by calculating the area under the curve.

- What's the probability that $0 < x < 10$?
- What's the probability that $x < 0$?
- What's the probability that $x > 10$?
- What's the probability that $x < 5$?

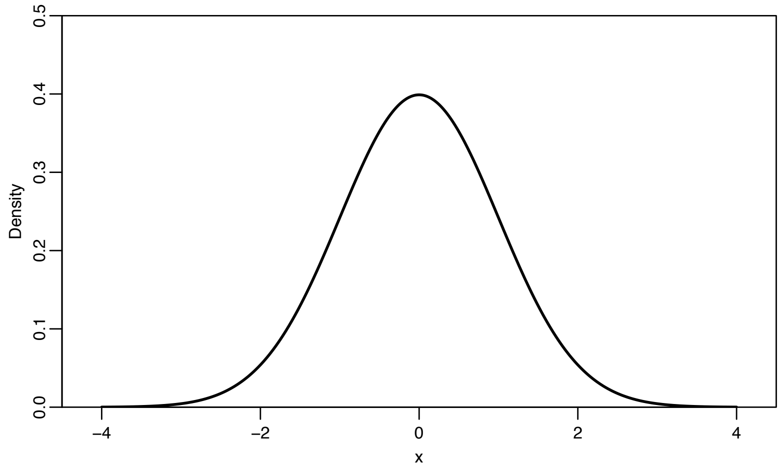
Normal Distribution

Let's say a random variable x is distributed Normally with mean 0 and standard deviation = 1

- Outcomes: any number between $-\infty$ and ∞.
- Once again, as there are infinite number of values, we cannot list up all values and associated probabilities.
- PDF for the Normal distribution with mean 0 and standard deviation = 1:



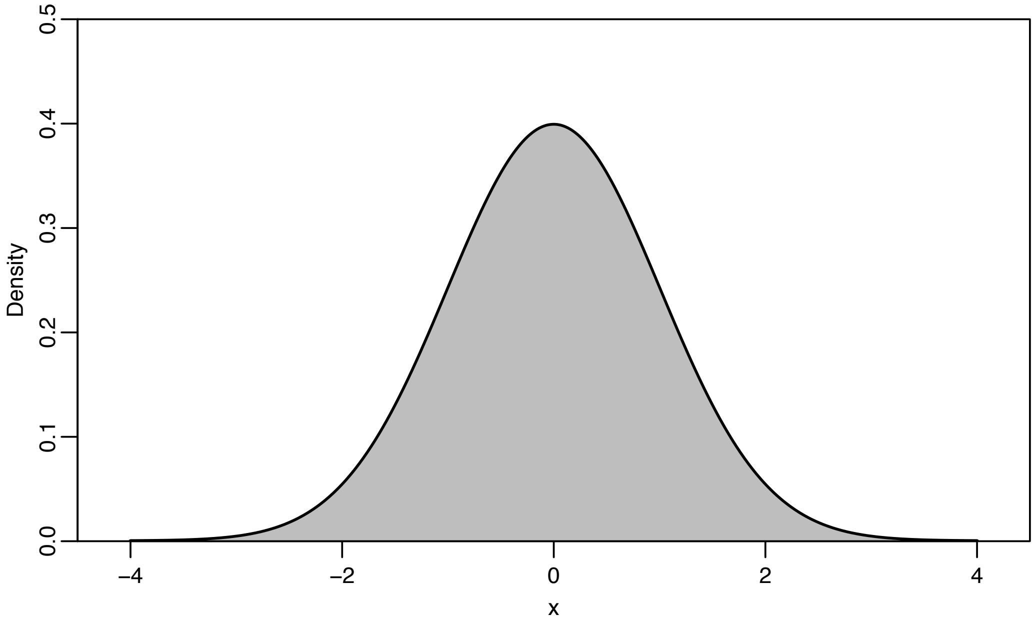
Normal Distribution



The area under the curve represents probabilities:

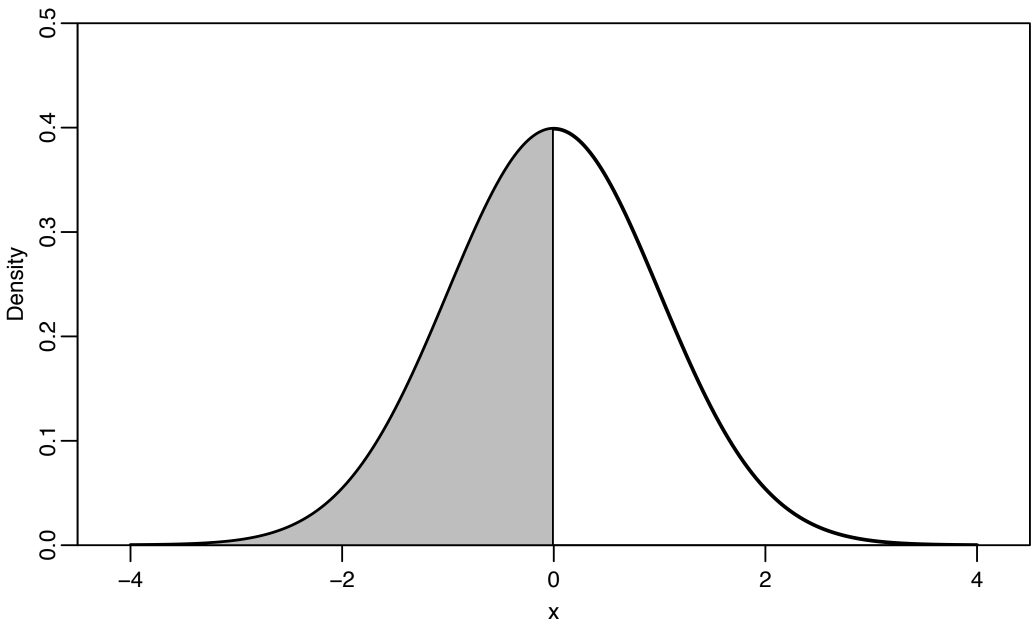
- What's the probability that $-\infty < x < \infty$?
- What's the probability that $x < 0$?
- What's the probability that $x > 0$?
- What's the probability that $-1 < x < 1$?
- What's the probability that $-2 < x < 2$?
- What's the probability that $-3 < x < 3$?

Normal Distribution



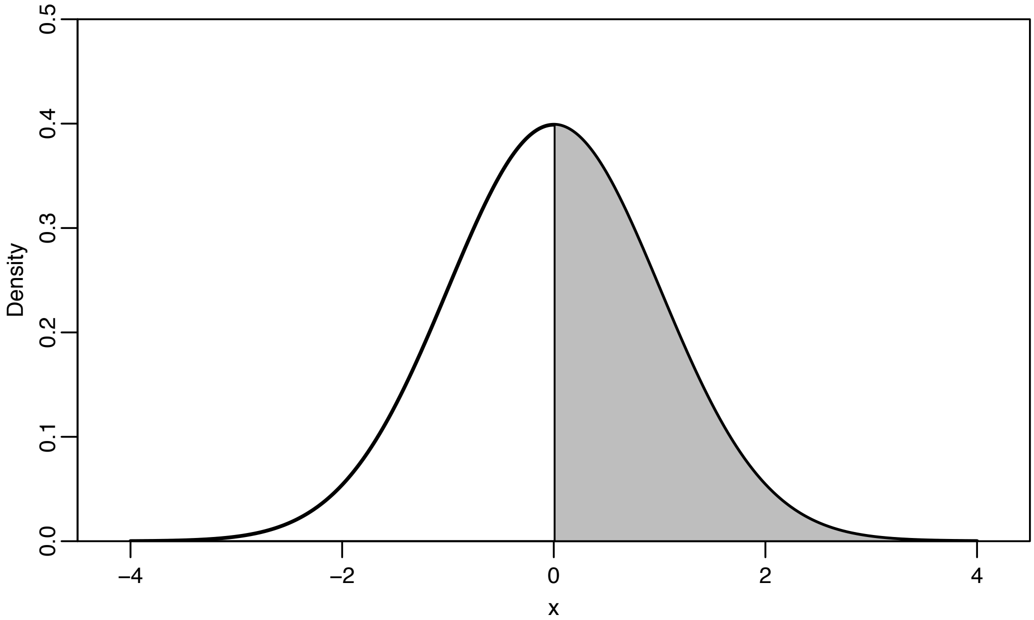
The probability that $-\infty < x < \infty$ is 1.

Normal Distribution



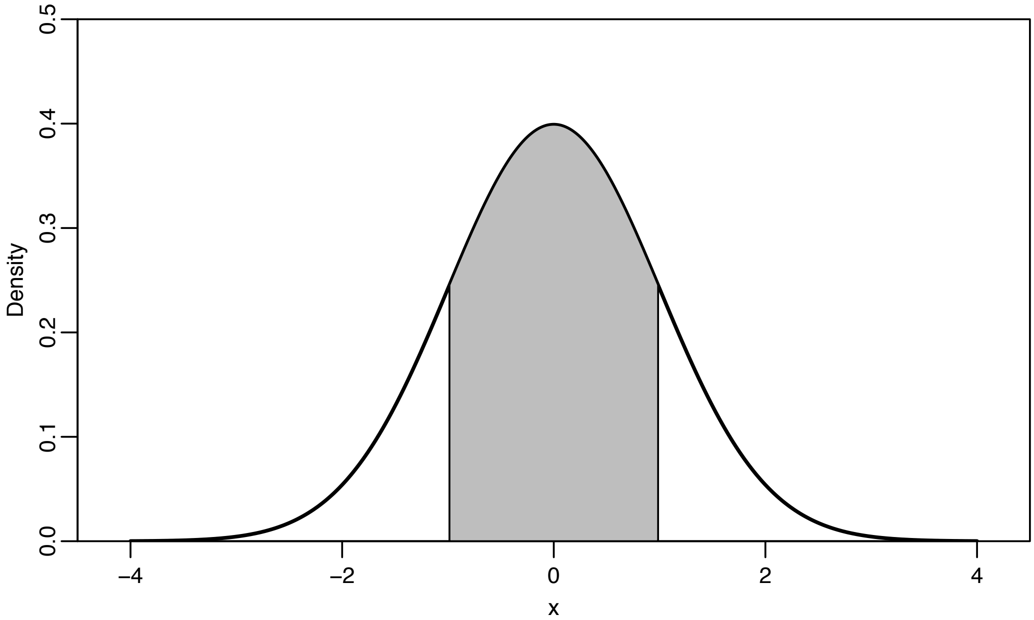
The probability that $x < 0$ is 0.5.

Normal Distribution



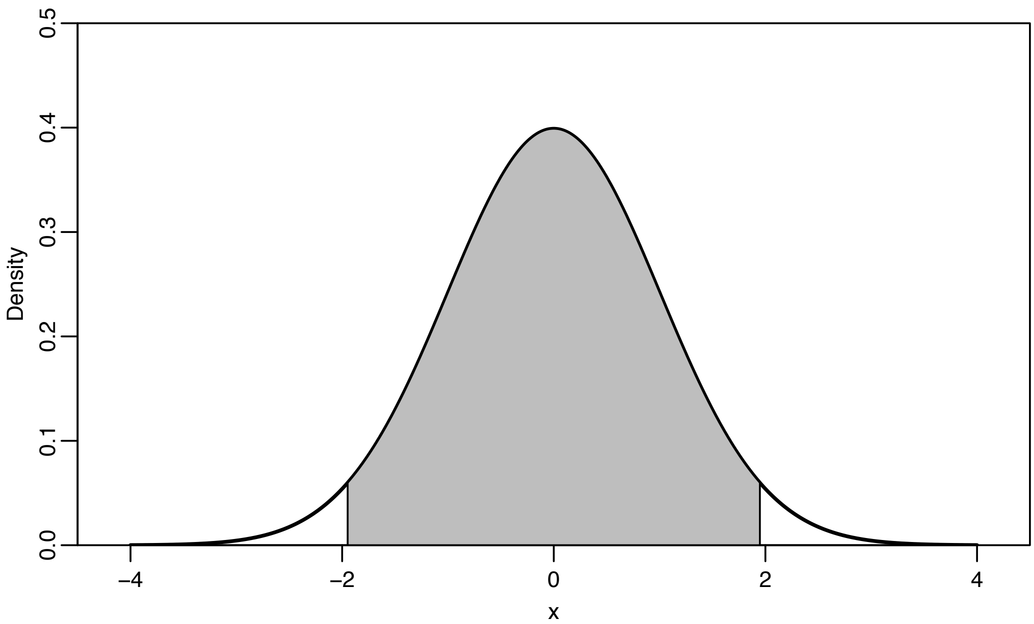
The probability that $x > 0$ is also 0.5.

Normal Distribution



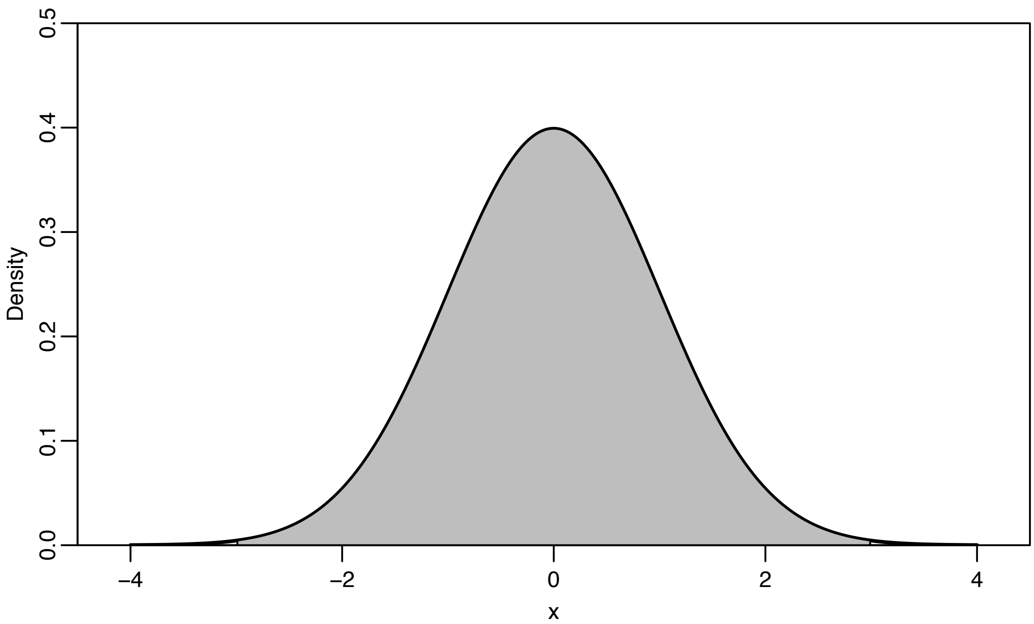
The probability that $-1 < x < 1$ is approximately 0.68.

Normal Distribution



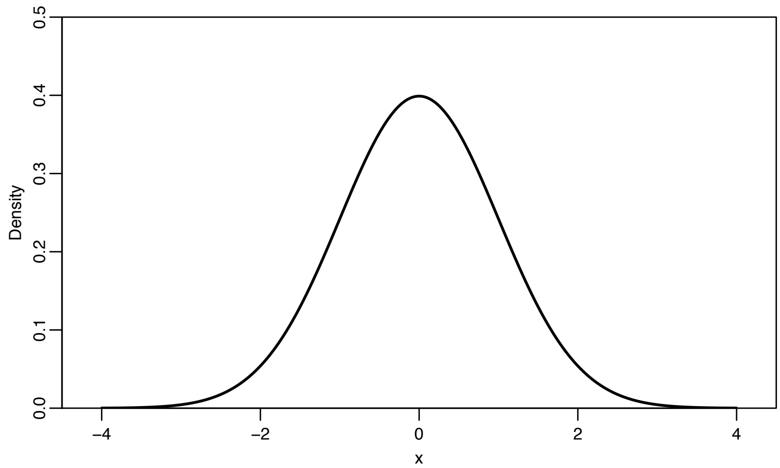
The probability that $-2 < x < 2$ is approximately 0.95.

Normal Distribution



The probability that $-3 < x < 3$ is approximately 0.99.

Normal Distribution



We can also ask the questions the other way around:

- What's the interval of x that makes 'the probability of observing a value as extreme (i.e., further away from 0) as them is equal to 0.05?

χ^2 Distribution

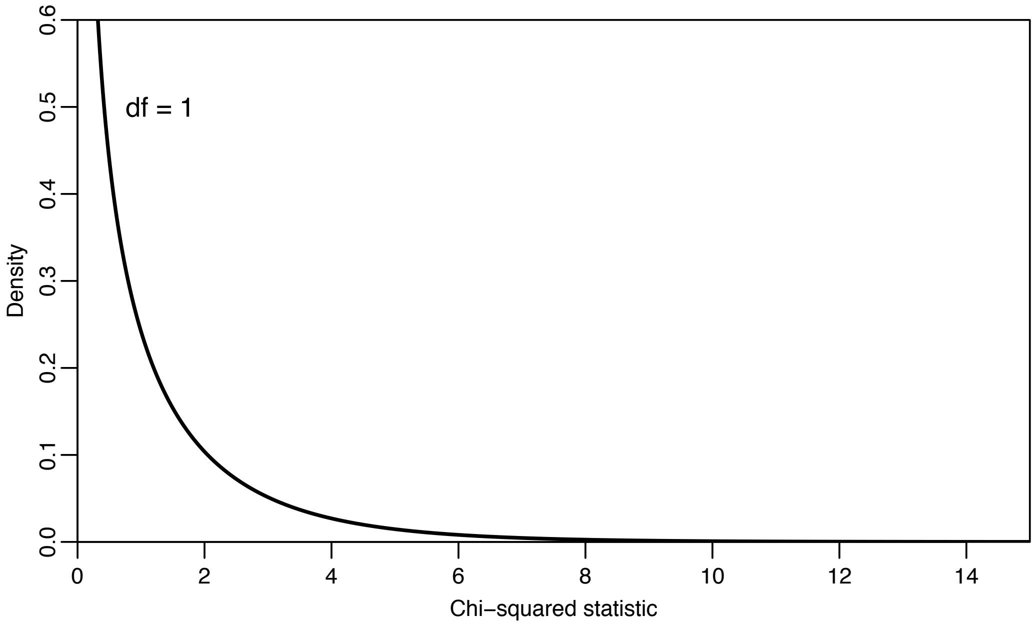
Recall:

- χ^2 statistic is always positive
- $\chi^2 = 0$ if the null hypothesis is true
- The shape of χ^2 distribution depends on the degree of freedom (df)

χ^2 Distribution

Recall:

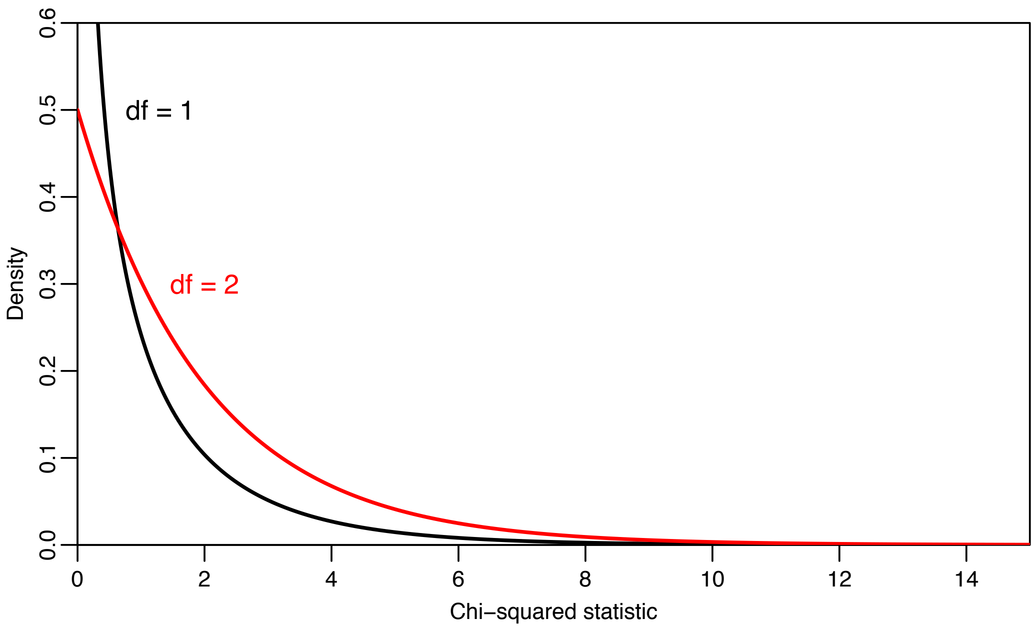
- χ^2 statistic is always positive
- $\chi^2 = 0$ if the null hypothesis is true
- The shape of χ^2 distribution depends on the degree of freedom (df)



χ^2 Distribution

Recall:

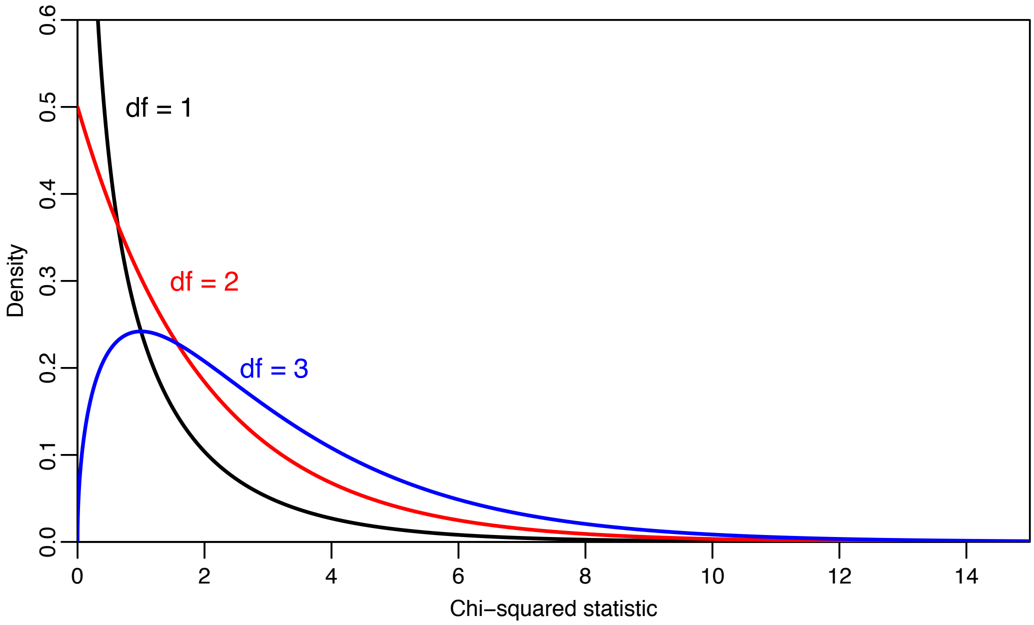
- χ^2 statistic is always positive
- $\chi^2 = 0$ if the null hypothesis is true
- The shape of χ^2 distribution depends on the degree of freedom (df)



χ^2 Distribution

Recall:

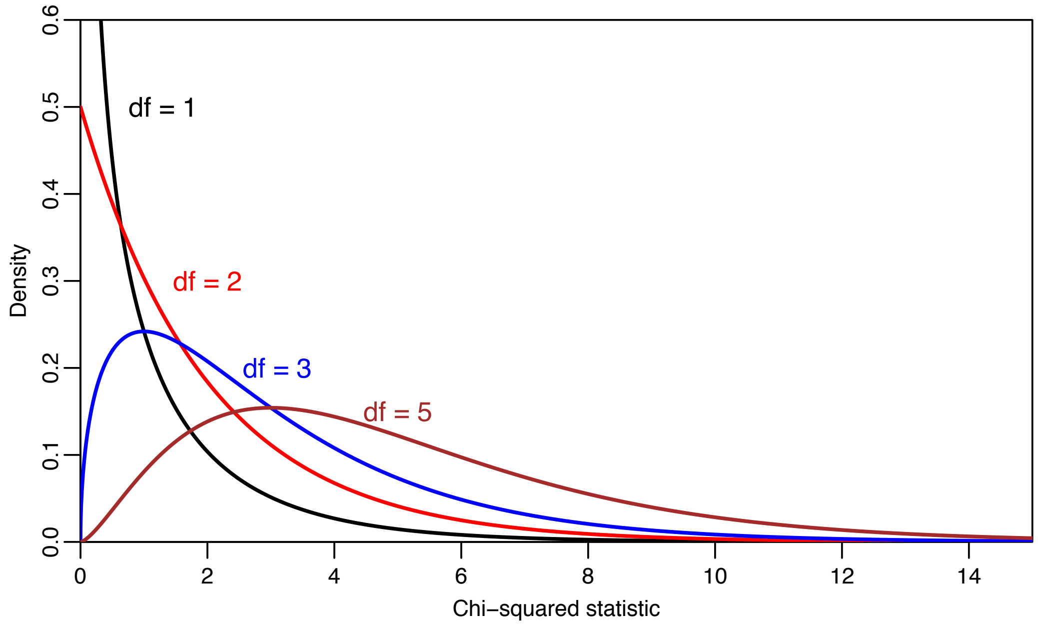
- χ^2 statistic is always positive
- $\chi^2 = 0$ if the null hypothesis is true
- The shape of χ^2 distribution depends on the degree of freedom (df)



χ^2 Distribution

Recall:

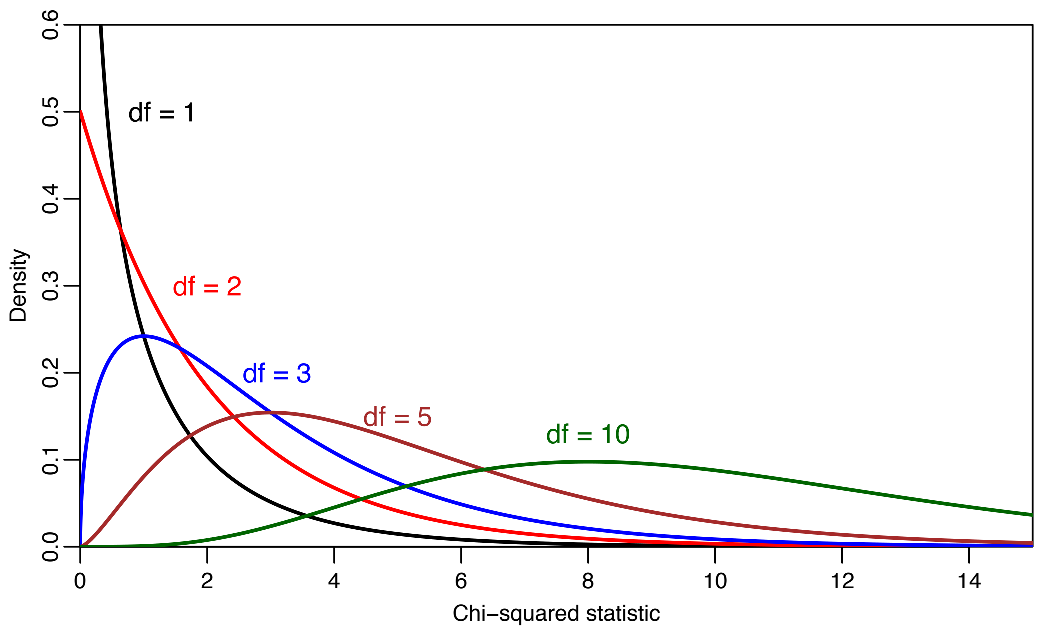
- χ^2 statistic is always positive
- $\chi^2 = 0$ if the null hypothesis is true
- The shape of χ^2 distribution depends on the degree of freedom (df)



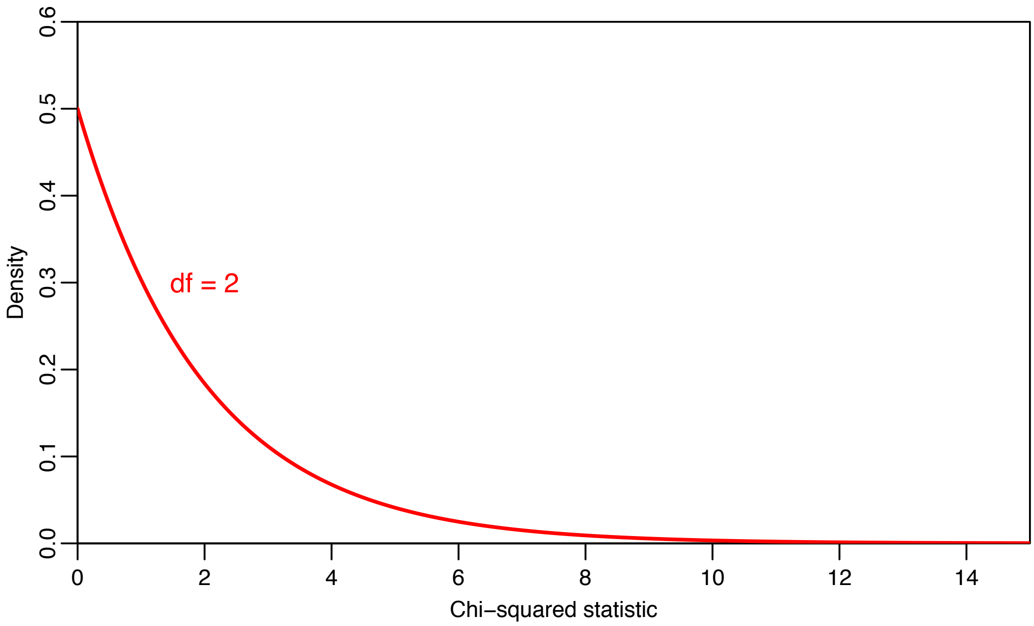
χ^2 Distribution

Recall:

- χ^2 statistic is always positive
- $\chi^2 = 0$ if the null hypothesis is true
- The shape of χ^2 distribution depends on the degree of freedom (df)

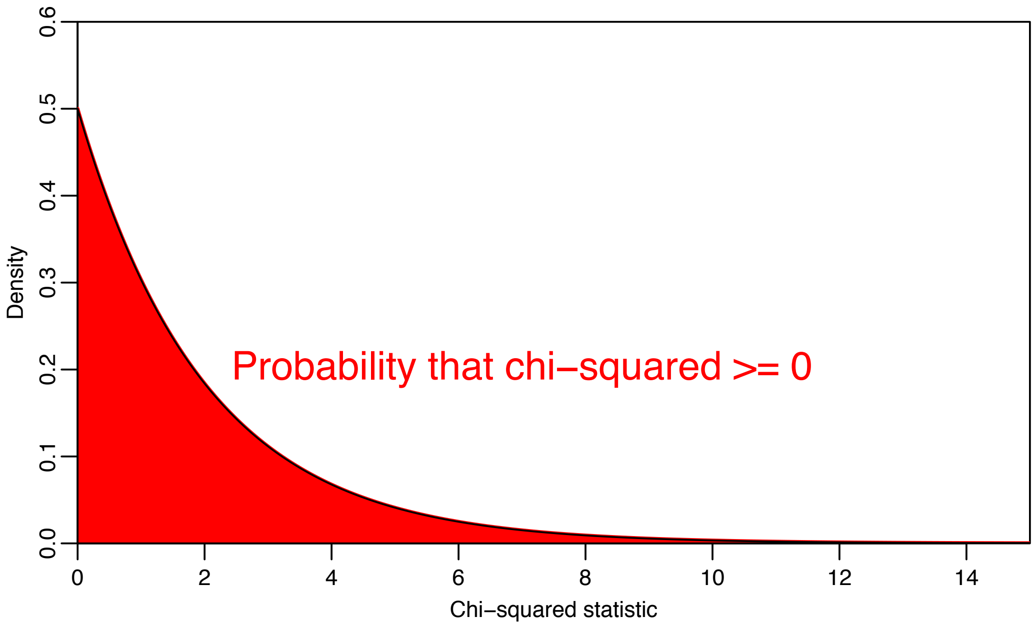


Area Under the Curve = Probability



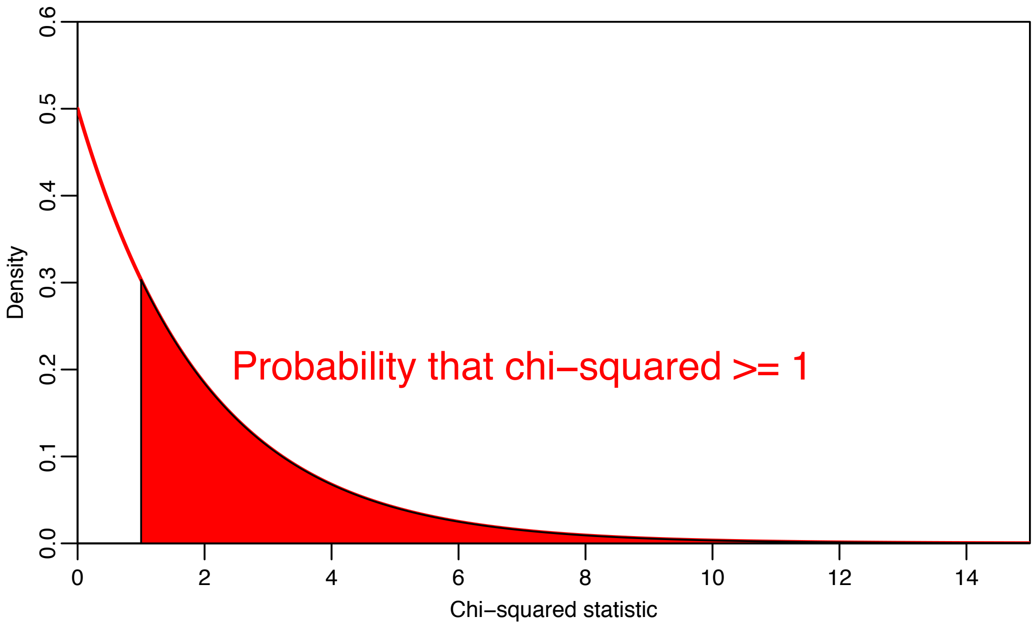
χ^2 statistic is always $\geq 0 \rightarrow$ probability that $\chi^2 \geq 0$ is ___.

Area Under the Curve = Probability



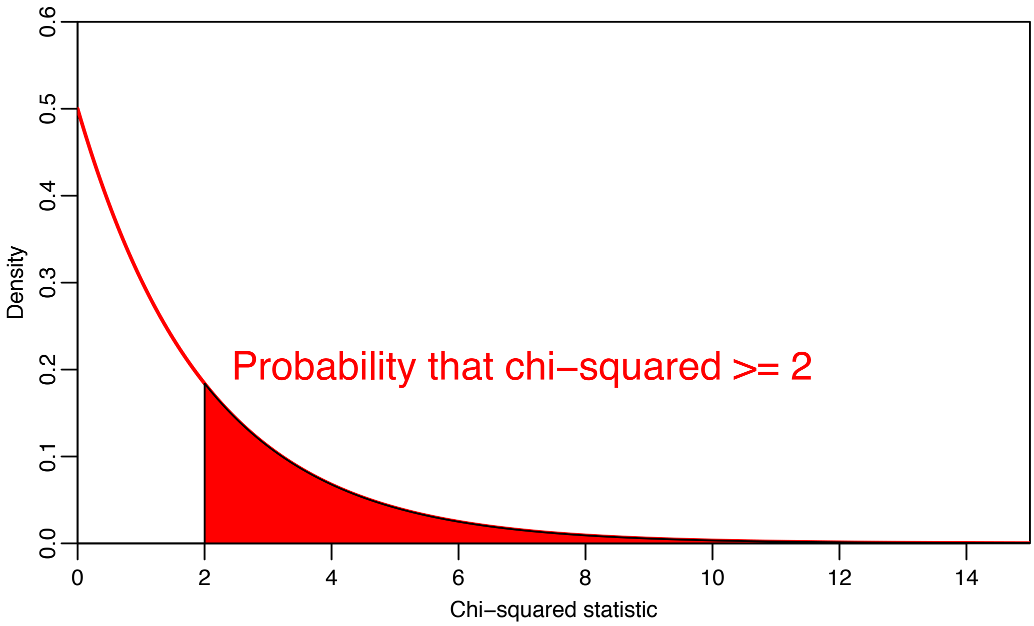
Area under the curve for $\chi^2 \geq 0$ represents the probability that $\chi^2 \geq 0$

Area Under the Curve = Probability



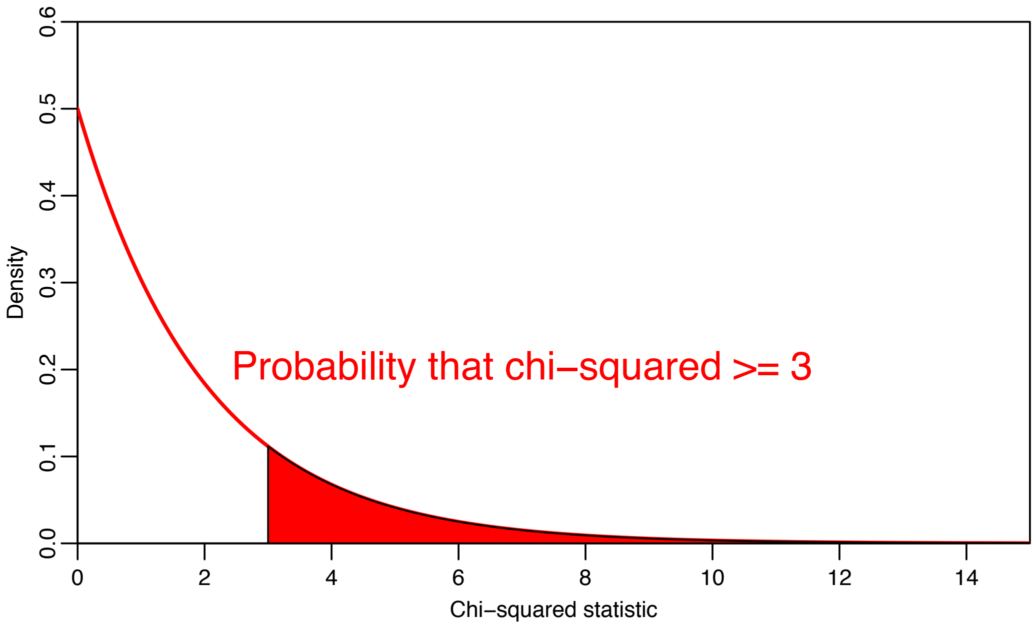
Area under the curve for $\chi^2 \geq 1$ represents the probability that $\chi^2 \geq 1$

Area Under the Curve = Probability



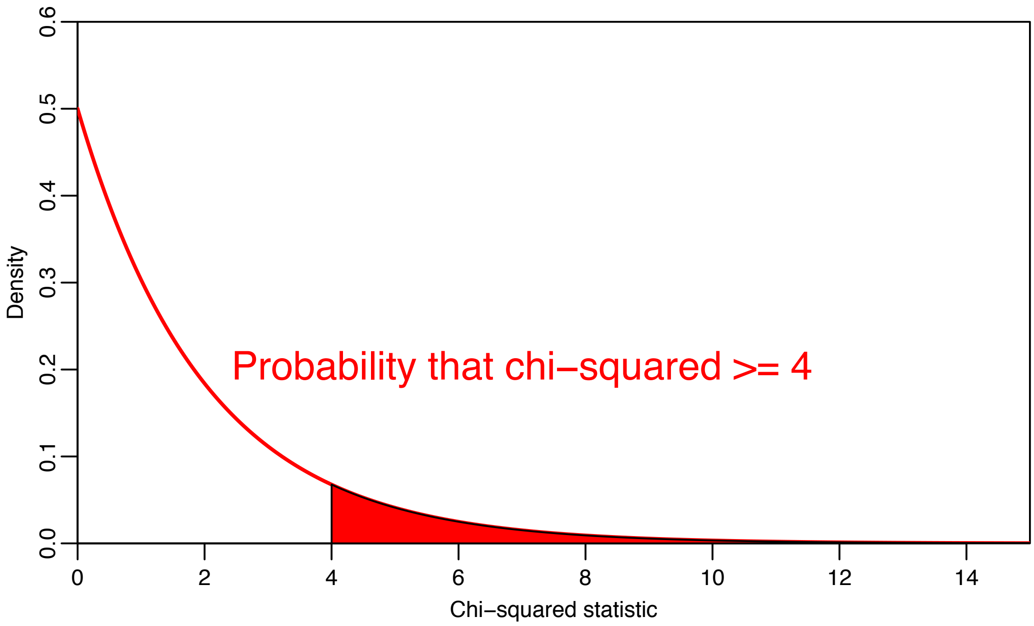
Area under the curve for $\chi^2 \geq 2$ represents the probability that $\chi^2 \geq 2$

Area Under the Curve = Probability



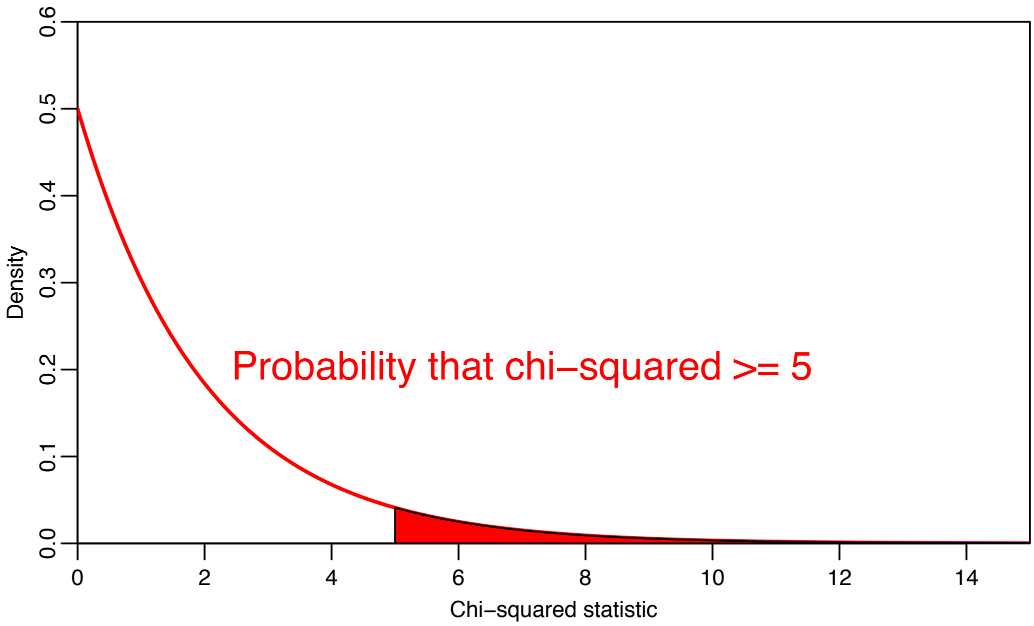
Area under the curve for $\chi^2 \geq 3$ represents the probability that $\chi^2 \geq 3$

Area Under the Curve = Probability



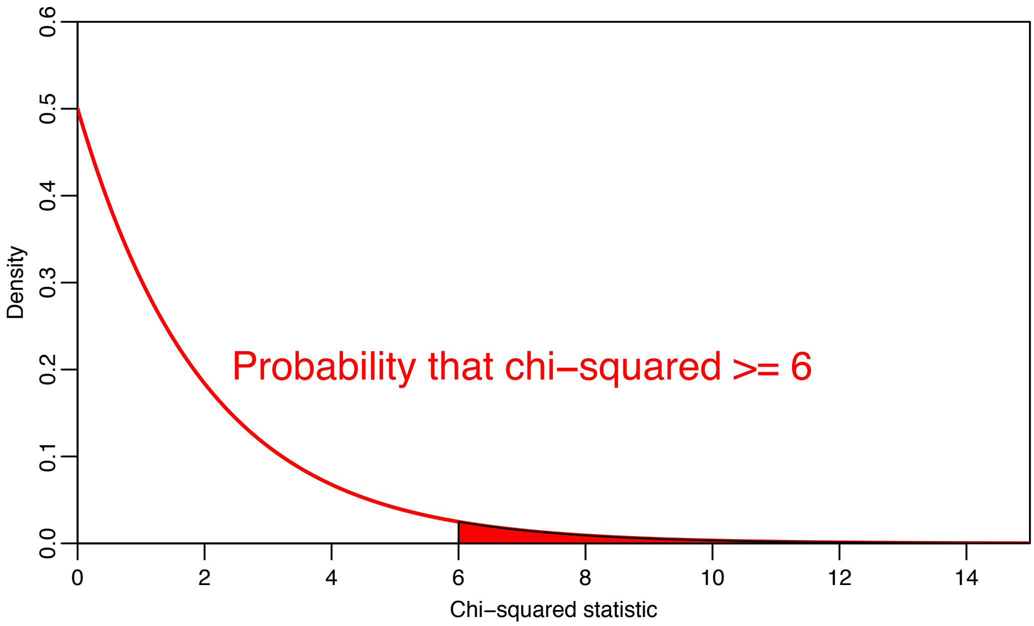
Area under the curve for $\chi^2 \geq 4$ represents the probability that $\chi^2 \geq 4$

Area Under the Curve = Probability



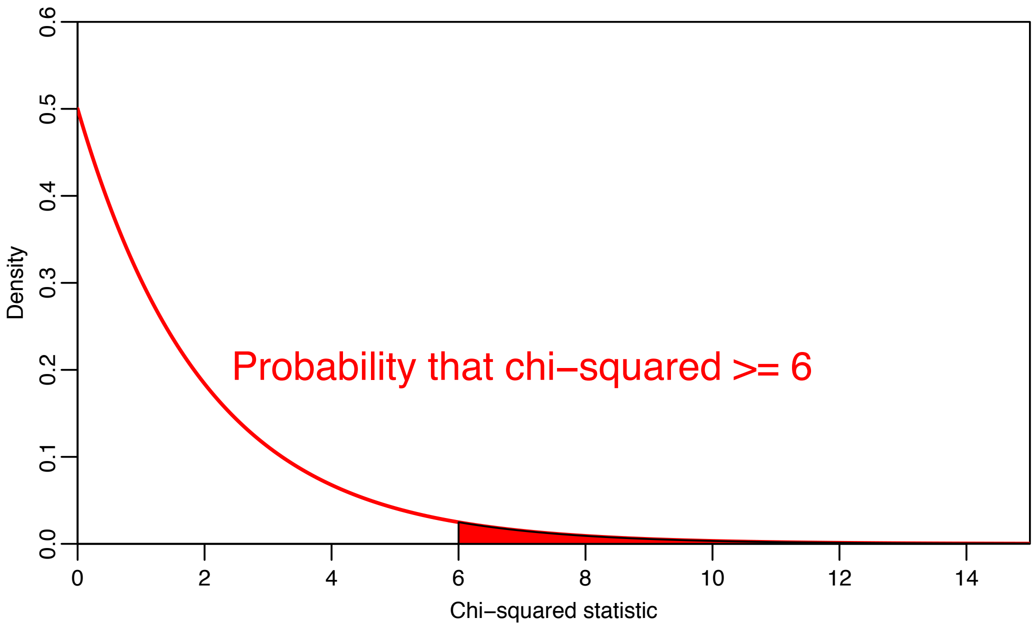
Area under the curve for $\chi^2 \geq 5$ represents the probability that $\chi^2 \geq 5$

Area Under the Curve = Probability



Area under the curve for $\chi^2 \geq 6$ represents the probability that $\chi^2 \geq 6$

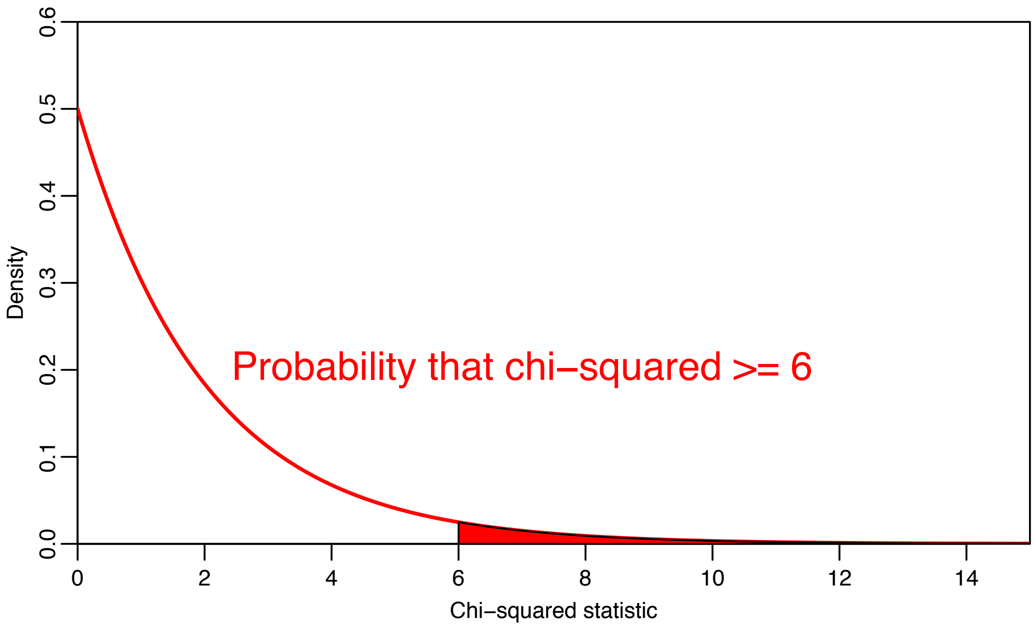
Area Under the Curve = Probability



Area under the curve for $\chi^2 \geq c$

- $=$ probability that $\chi^2 \geq c$
- $=$ *p*-value we obtain when the χ^2 statistic is equal to c

Area Under the Curve = Probability



Statistical tables (e.g., page 295)

- tell us the values of χ^2 at which *p*-value is equal to certain values
- tell us the range of *p*-values for given values of χ^2 we obtain

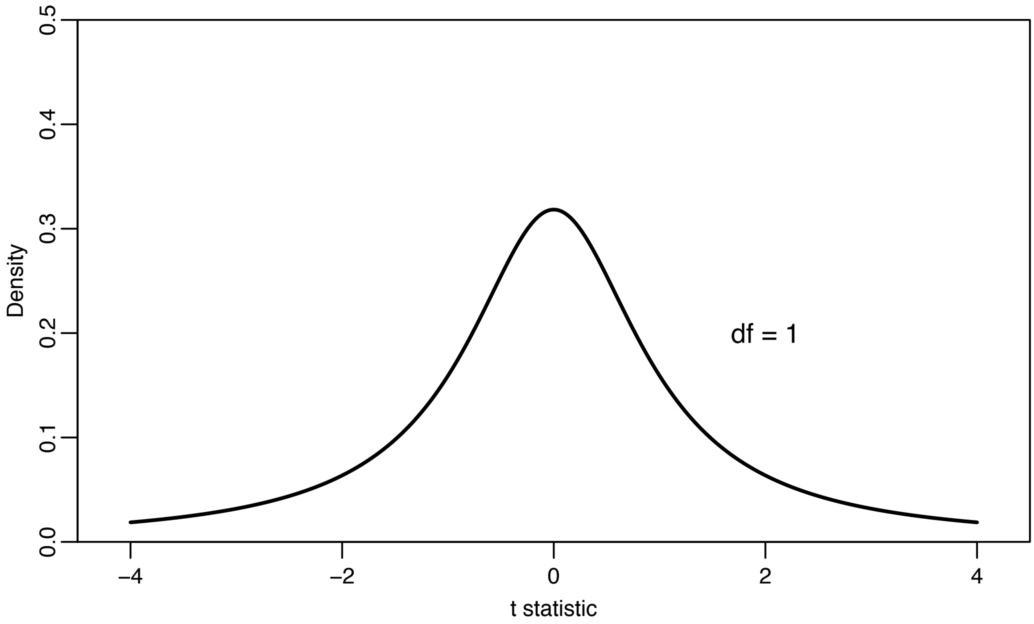
t-Distribution

Let's go back to the example: we obtained t statistic $= -5.145$.

- Unlike χ^2 statistic, t-statistic can take both negative and positive values.
- Just like χ^2 distribution, the shape of t distribution depends on the degree of freedom (df).
- When df is greater than 30, it is almost indistinguishable from Normal distribution with mean 0 and s=1.

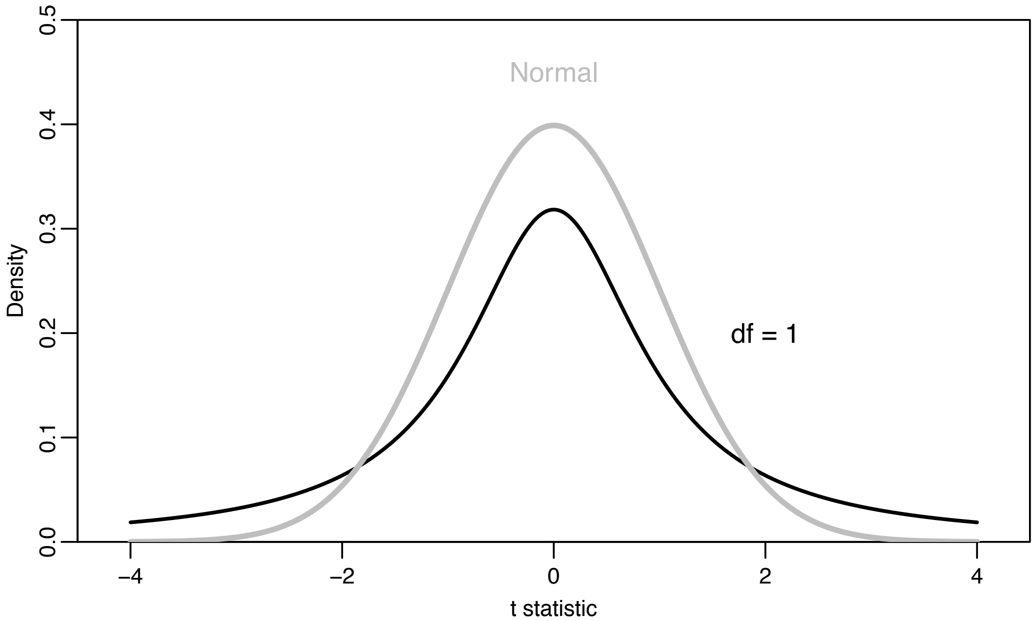
t-Distribution

t distribution with $df = 1$



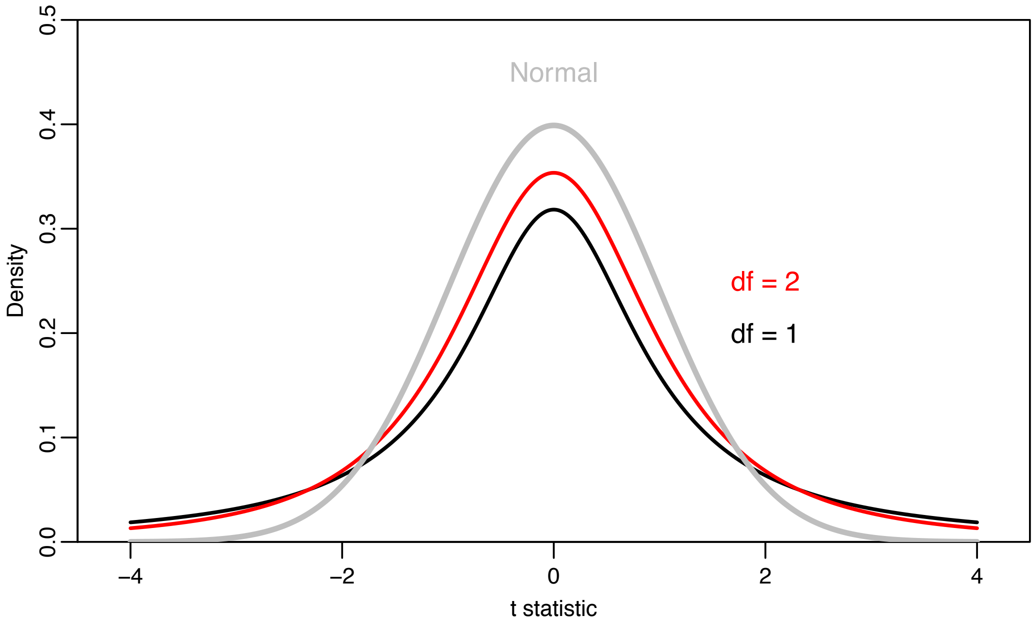
t-Distribution

t distribution with $df = 1$ with Normal (in gray)



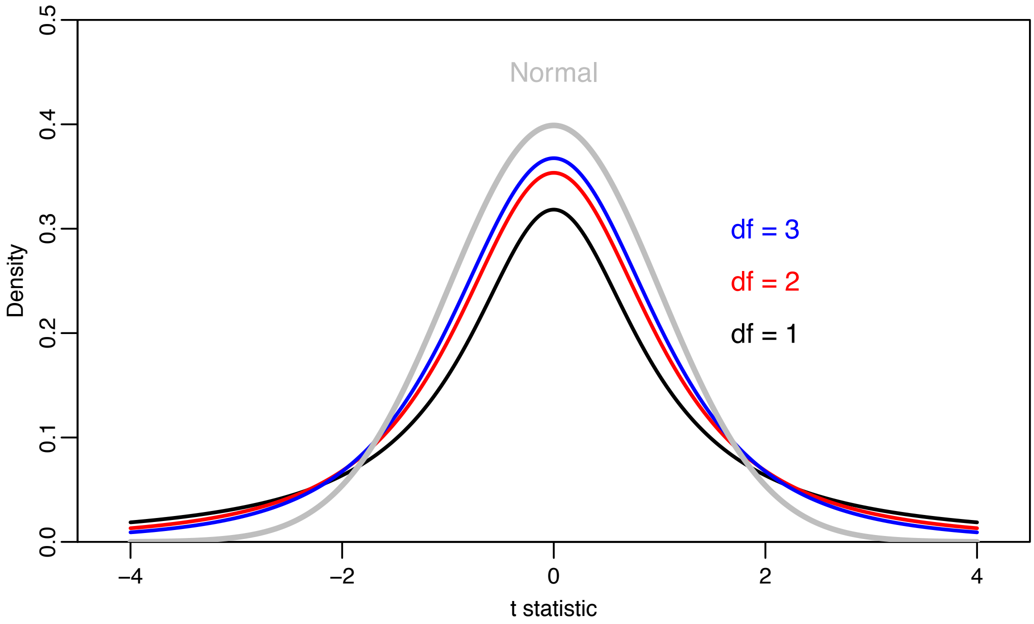
t-Distribution

t distribution with $df = 2$



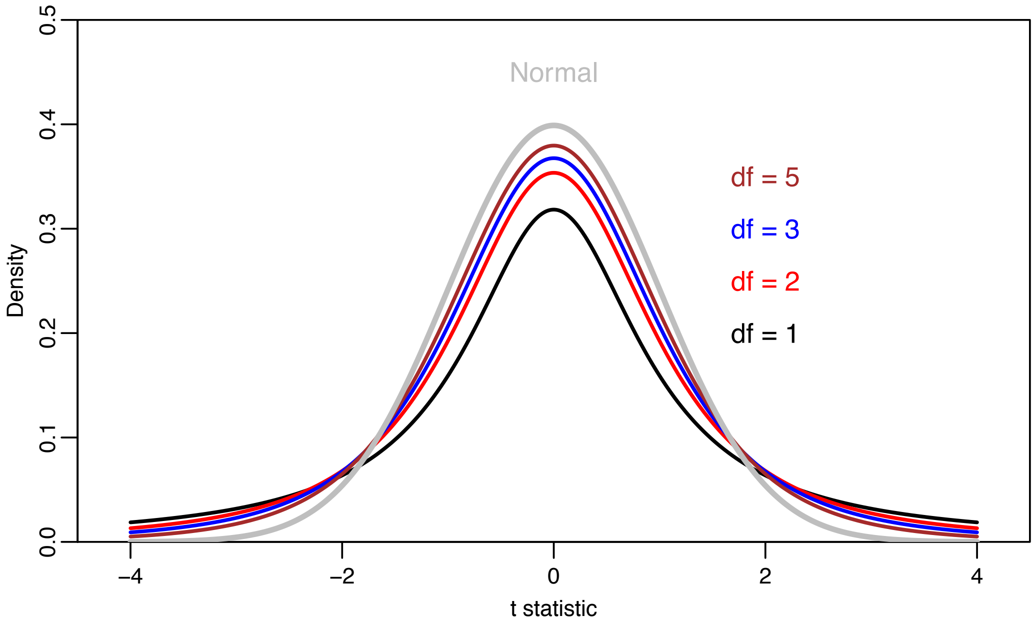
t-Distribution

t distribution with $df = 3$



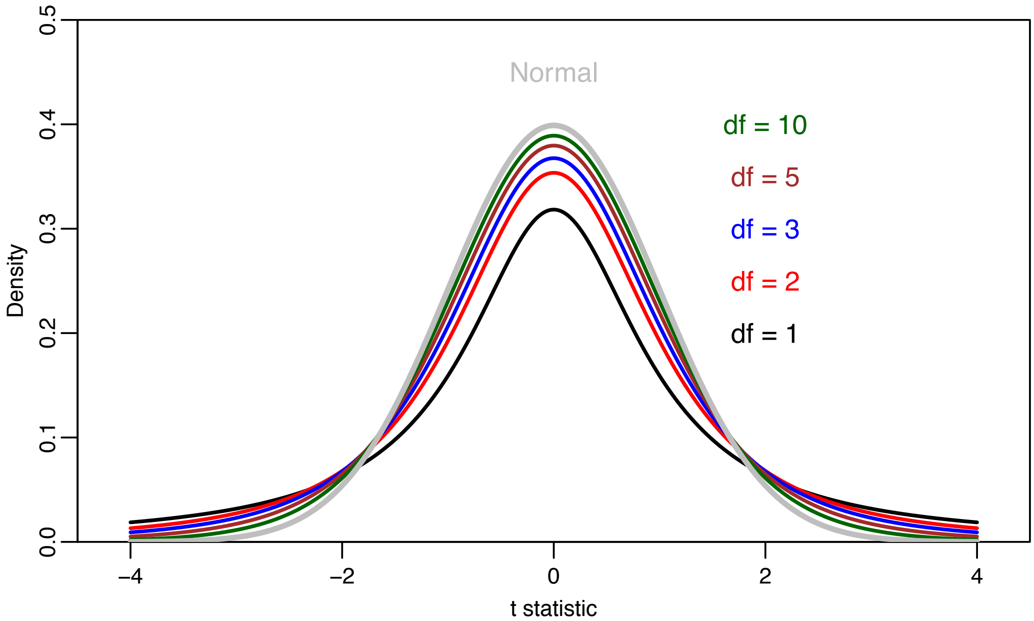
t-Distribution

t distribution with $df = 5$



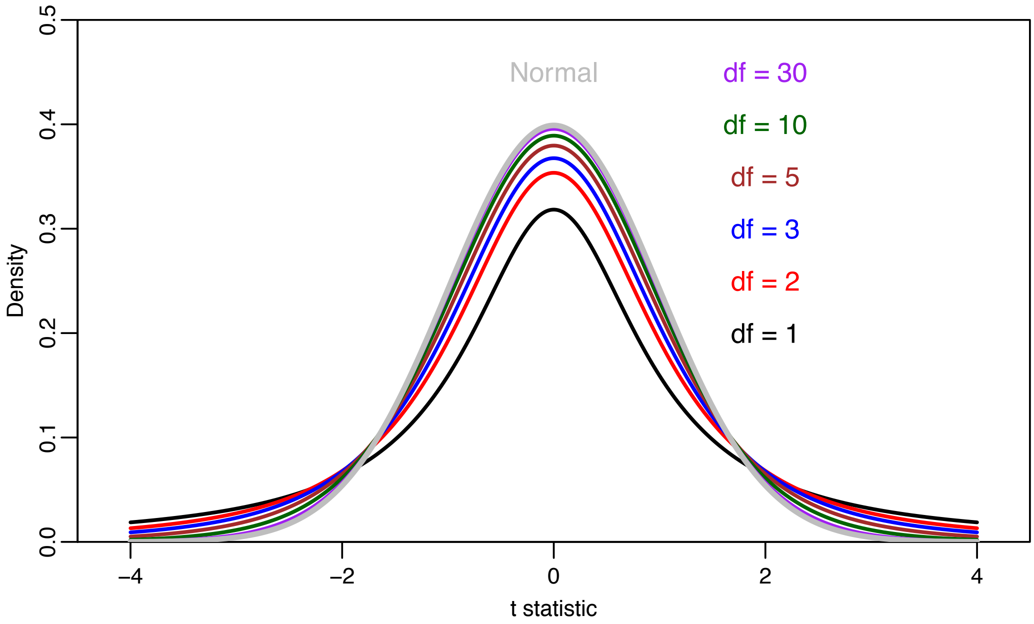
t-Distribution

t distribution with $df = 10$



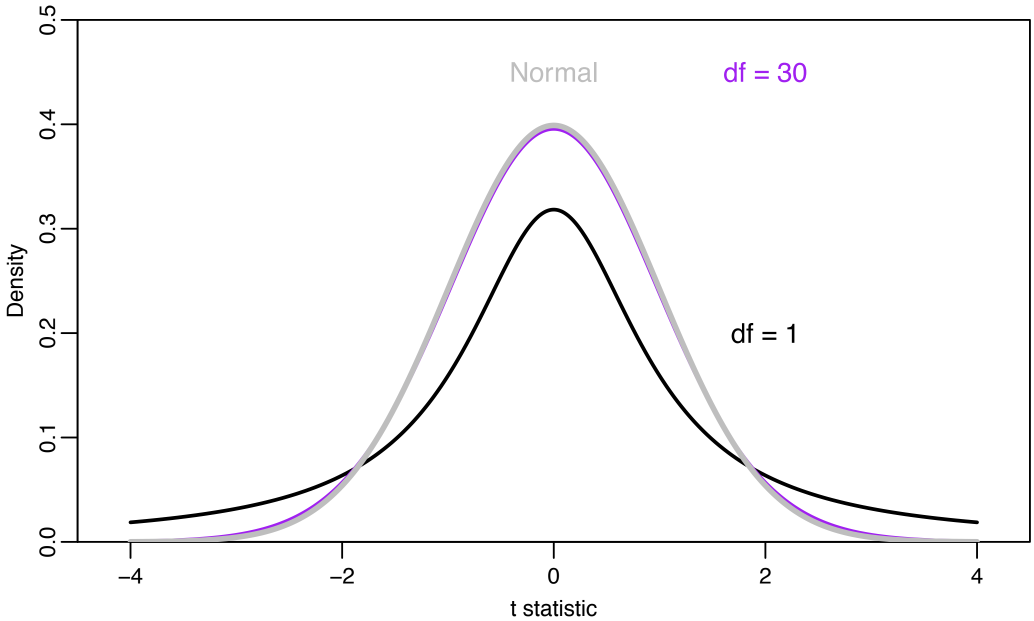
t-Distribution

t distribution with $df = 30$



t-Distribution

t distribution with $df = 30$



t-Statistic

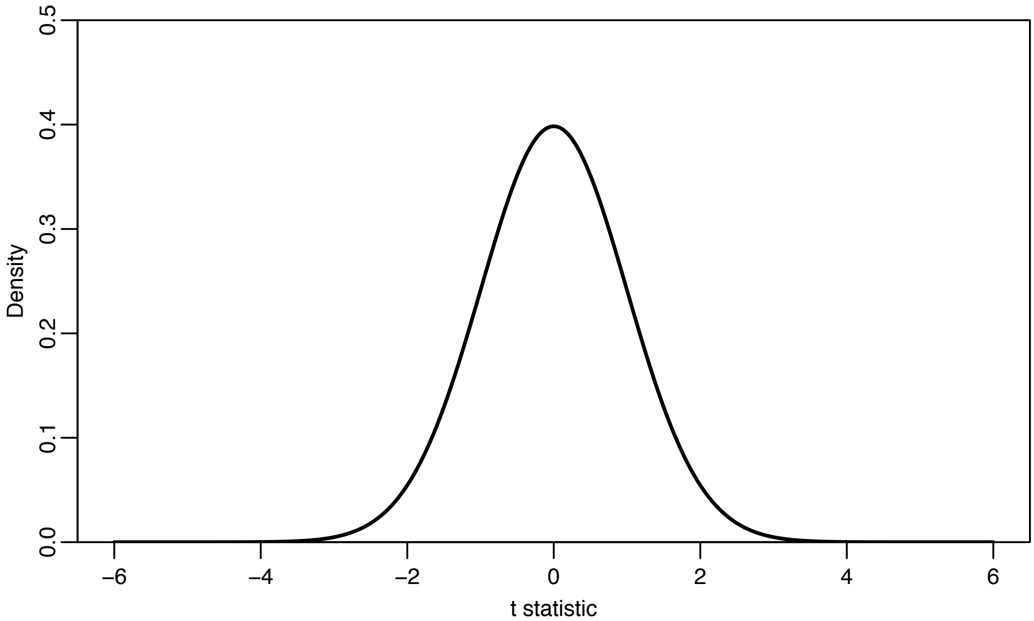
We obtained t statistic $= -5.145$.

Degree of freedom for a difference of means is given as

$$n_1 + n_2 - 2 = 114 + 66 - 2 = 178$$

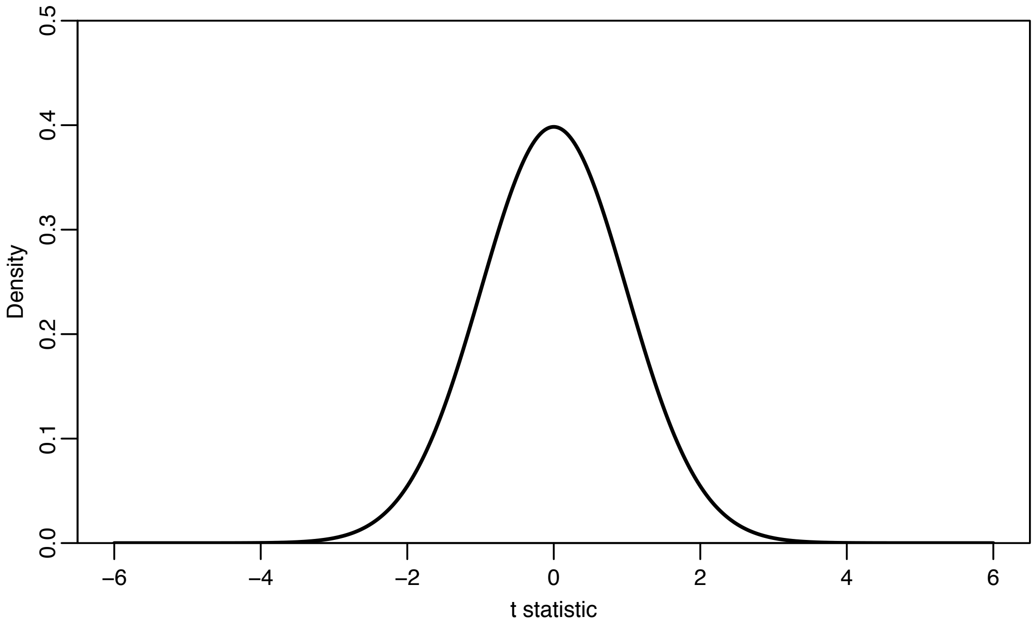
t-Statistic

- t-distribution with degree of freedom = 178
- t-statistic is 0 if the null is true.



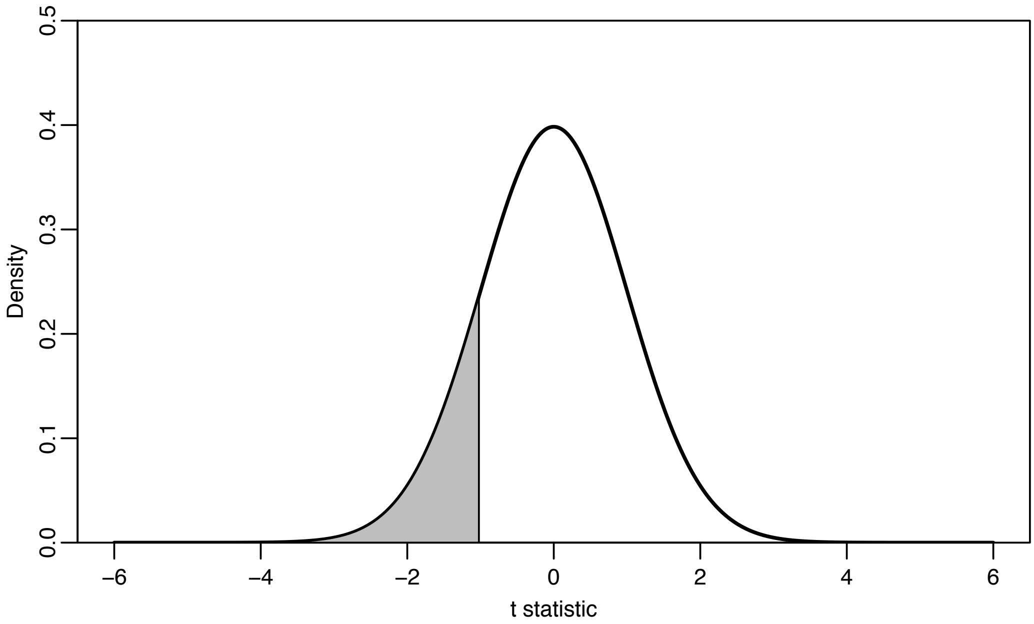
Quiz

How do we calculate the probability that $t < -1$?
(How do we calculate the p-value when t statistic is -1?)



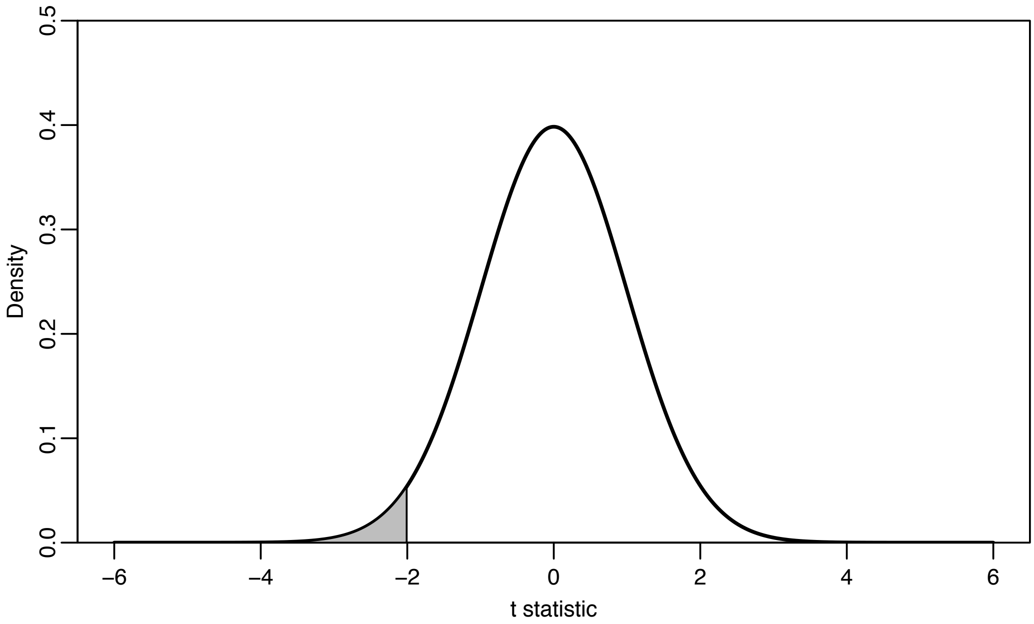
Quiz

Area under the curve for $t < -1$ gives the probability that $t < -1$. This is the p-value when t statistic is -1.



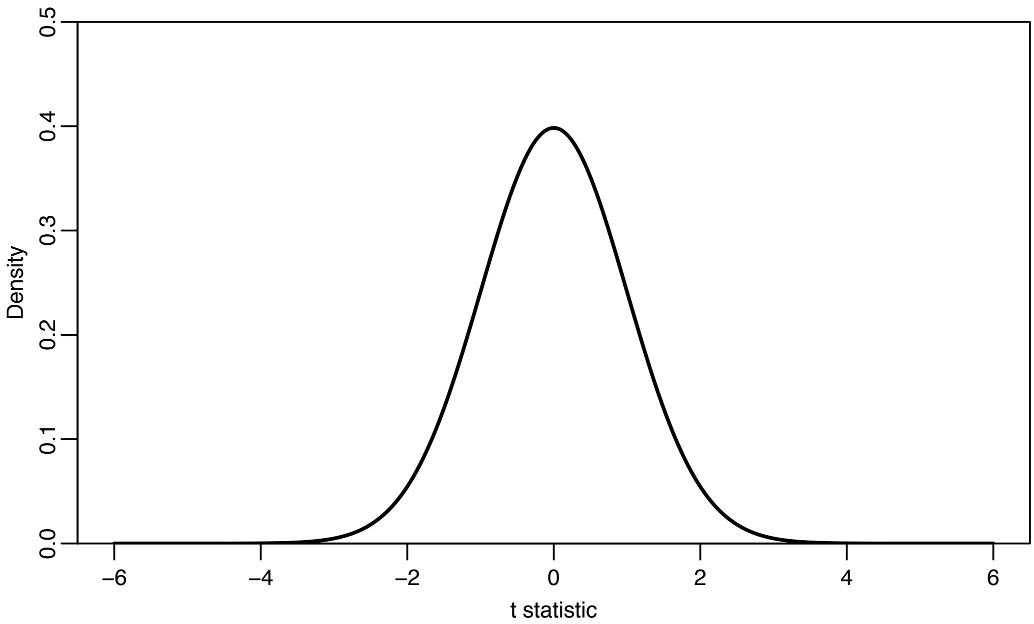
Quiz

Area under the curve for $t < -2$ gives the probability that $t < -2$. This is the p-value when t statistic is -2.



t-Statistic was -5.145

p-value is obtained by calculating the area under curve for $t < -5.145$



You can't really see in the graph, as it's so tiny.

Reading Statistical Tables

- Another way to obtain *p*-value is to refer to the statistical table on page 296.
- A bit tricky, as the cell entries are represented in absolute values.
- For example, the condition $t < -2$ means the absolute value of $t > 2$.

Reading Statistical Tables

	Level of significance					
df	**0.10**	**0.05**	**0.025**	**0.01**	**.005**	**0.001**
1	3.078	6.314	12.706	31.821	63.657	318.313
2	1.886	2.920	4.303	6.965	9.925	22.327
3	1.638	2.353	3.182	4.541	5.841	10.215
			...			
80	1.292	1.664	1.990	2.374	2.639	3.195
90	1.291	1.662	1.987	2.368	2.632	3.183
100	1.290	1.660	1.984	2.364	2.626	3.174
∞	1.282	1.645	1.960	2.326	2.576	3.090

Let's say our t-statistics were -2 (with df $= 100$)

- What is the minimum p-value we can get?
- What is the maximum level of confidence we can get?

Reading Statistical Tables

	Level of significance					
df	**0.10**	**0.05**	**0.025**	**0.01**	**.005**	**0.001**
1	3.078	6.314	12.706	31.821	63.657	318.313
2	1.886	2.920	4.303	6.965	9.925	22.327
3	1.638	2.353	3.182	4.541	5.841	10.215
			...			
80	1.292	1.664	1.990	2.374	2.639	3.195
90	1.291	1.662	1.987	2.368	2.632	3.183
100	1.290	1.660	1.984	2.364	2.626	3.174
∞	1.282	1.645	1.960	2.326	2.576	3.090

As our t-statistics is -5.145 (with df $= 178$)

- What is the minimum p-value we can get?
- What is the maximum level of confidence we can get?

Reading Stata Output

```
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
      No |     114    14.15965    .8859928    9.459815    12.40434    15.91496
     Yes |      66    22.38939    1.442365    11.71783    19.50879    25.26999
---------+--------------------------------------------------------------------
combined |     180    17.17722    .8238416    11.05299    15.55153    18.80291
---------+--------------------------------------------------------------------
    diff |           -8.229745    1.599568               -11.3863   -5.073188
------------------------------------------------------------------------------
    diff = mean(No) - mean(Yes)                                   t =  -5.1450
Ho: diff = 0                                     degrees of freedom =      178

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 1.0000
```

- Our causal theory expects `diff` to be negative.
- t-statistic is -5.1450, which yields a p-value < 0.0001.
- We reject the null hypothesis that there is no difference in mean levels of female representation between PR system and majority system.

Bivariate Hypothesis Test 3: Correlation Analysis

- Both Y and X are continuous.

- We follow the same logic and same steps as cross-tabulation analysis and difference of means test:
  1. Form the null and alternative
  2. Examine and describe the sample
  3. Compare the observed and expected
  4. Reject or not reject the null

Labour Rights Protection and Unionisation

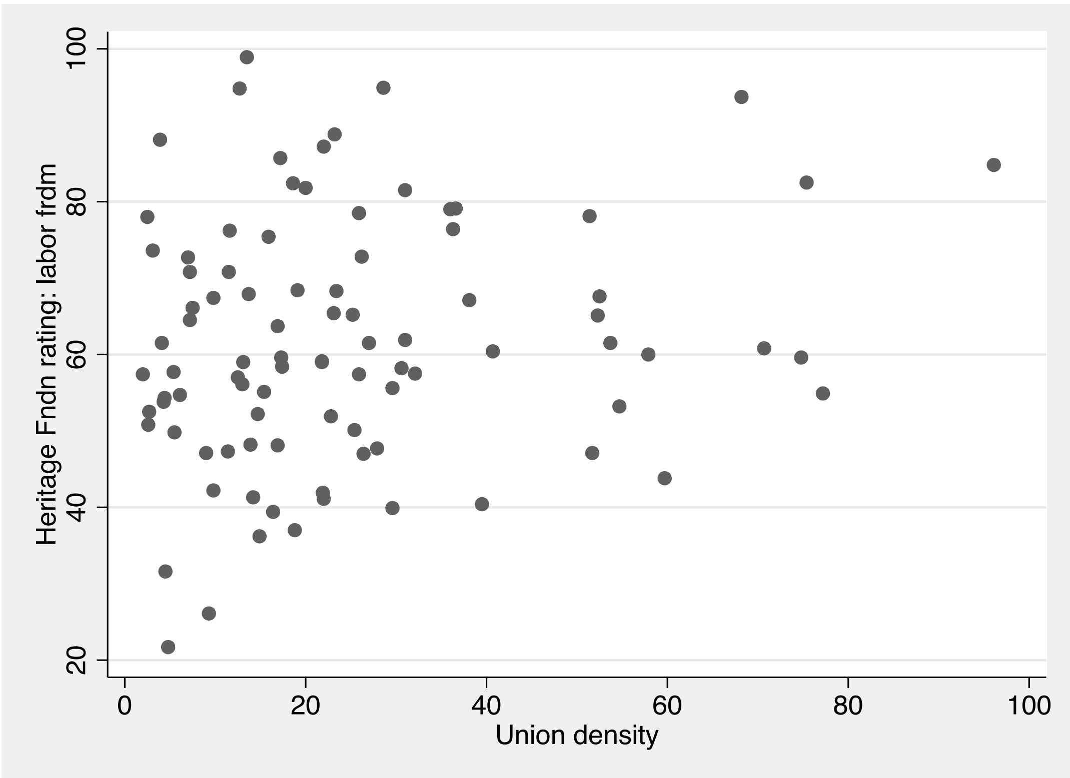
As more labourers join unions, unions can exert greater influence on the government. Therefore, the greater the level of unionisation, the better the protection of labourer rights.

Hypothesis

There will be a positive relation between levels of unionisation and labour rights protection.

- Y: Labour rights protection (0–100)
- X: Levels of unionisation (% of labourers in unions: 0–100)
- The null hypothesis: there is no relationship between the two

Describe the Relationship in the Sample: Do It Graphically



Describe the Relationship in the Sample: Do It Numerically

Covariance between X and Y is:

$$\mathrm{cov}_{XY} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n}$$

$(X_i - \bar{X})(Y_i - \bar{Y})$ is positive when

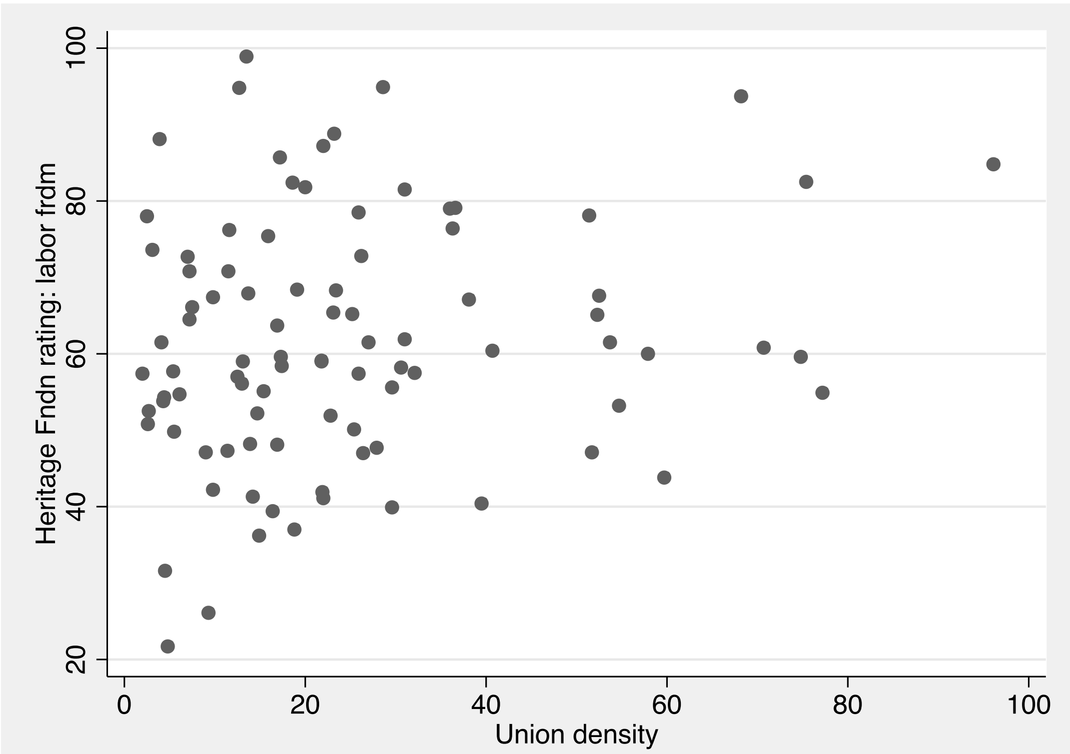
$(X_i - \bar{X})$ and $(Y_i - \bar{Y})$ are both positive or both negative.

$(X_i - \bar{X})(Y_i - \bar{Y})$ is negative when

$(X_i - \bar{X})$ and $(Y_i - \bar{Y})$ have different signs.

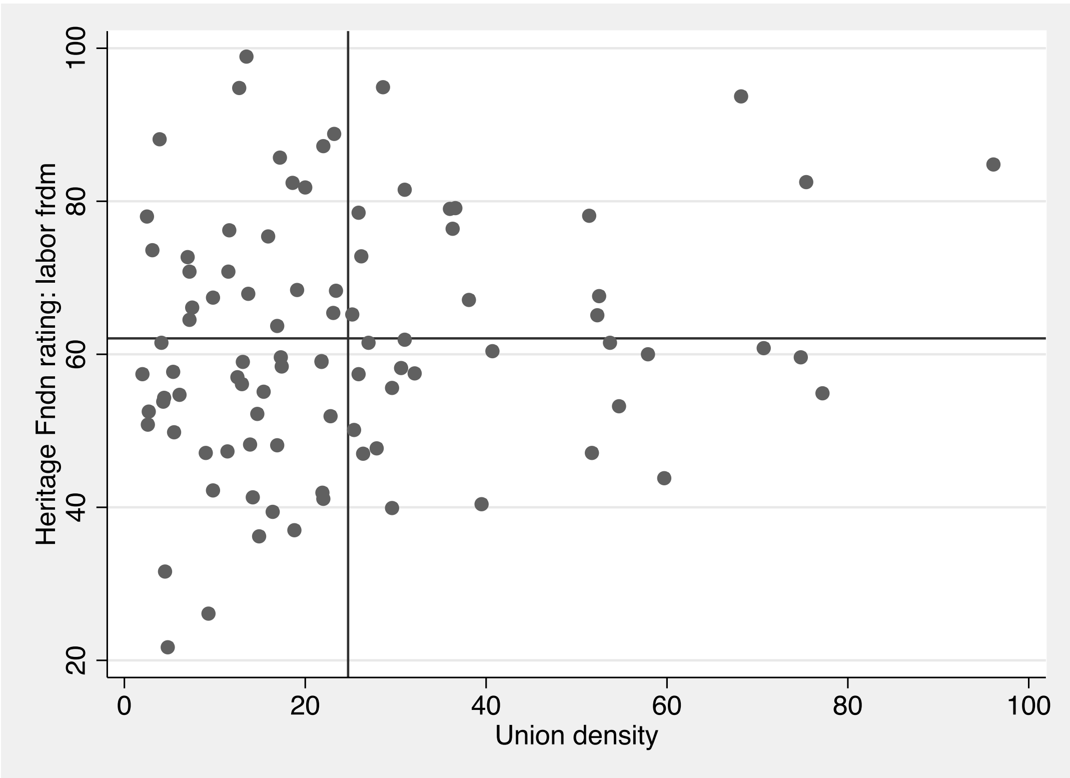
Covariance

$$\text{cov}_{XY} = \left(\sum (X_i - \bar{X})(Y_i - \bar{Y})\right)/n$$



Covariance

$$\text{cov}_{XY} = (\sum (X_i - \bar{X})(Y_i - \bar{Y}))/n$$



From Covariance to Correlation

Covariance

$$-\sqrt{s_X^2 \ s_Y^2} < \mathrm{cov}_{XY} < \sqrt{s_X^2 \ s_Y^2}$$

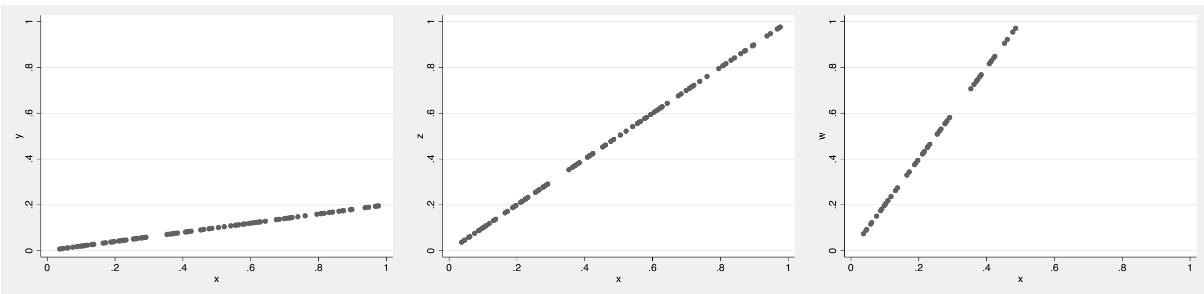
Correlation coefficient

$$r = \frac{\mathrm{cov}_{XY}}{\sqrt{s_X^2 \ s_Y^2}}$$

- $-1 < r < 1$
- t statistic for correlation is $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$ with $\mathrm{df} = n - 2$.
- p-value for correlation can also be obtained.

Notes on r

- r is equal to 1 (and p-value < 0.0001) when all the observations are aligned on a single line with a positive slope.
- However, greater values of r (or smaller values of p) do not suggest stronger relationship between X and Y.
- All of the following three samples generate $r = 1$.



Reading Stata Output

```
             | free_l~r   unions
-------------+------------------
  free_labor |   1.0000
             |
             |
      unions |   0.1781   1.0000
             |   0.0913
```

- $r = 0.1781$
- p-value is 0.0913
- The correlation is positive (as expected by our causal theory) and statistically significant at 10% significance level (We reject the null hypothesis at 90% confidence level).
- We fail to reject the null hypothesis at the conventional 95% confidence level.

Summary

- Bivariate hypothesis tests
  - Cross-tabular analysis (χ^2 test)
  - Difference of means analysis (t test)
  - Correlation analysis (t test)
- Steps
  1. Form the null and alternative hypotheses
  2. Describe the pattern: calculate some statistic
  3. Compare the obtained statistic with some threshold value
  4. When the obtained statistic is greater/smaller than the threshold, we reject the null
- Establishing a bivariate relationship is only the starting point!