MORU Tropical Health Network

Basic R programing

11-12 Jan 2021

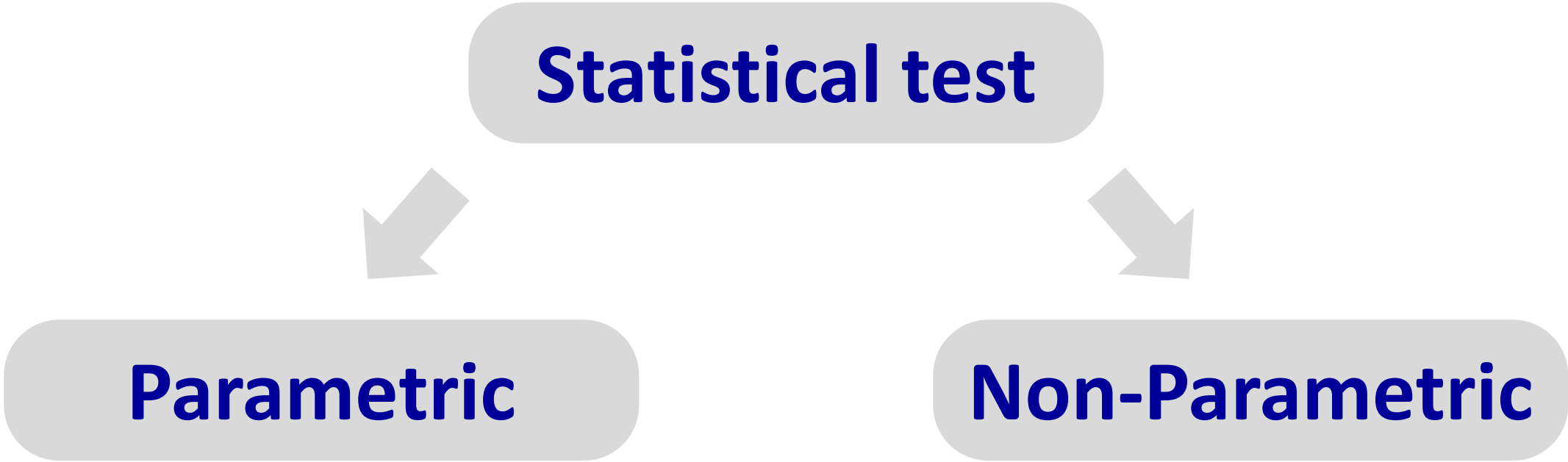
Lecture 4 (09:10-10:30): Probability and Hypothesis Testing

Dr. Palang Chotsiri

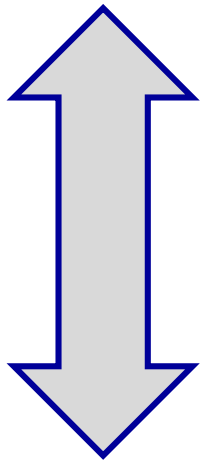
palang@tropmedres.ac

Learning outline:

Date-time	Outline
Day 2:	
09:00 – 09:10	Recap
09:10 – 10:30	Probability and Hypothesis testing
10:30 – 10:45	Break
10:45 – 12:00	Categorical data analysis
12:00 – 13:00	Lunch
13:00 – 14:30	Continuous data analysis (linear modelling)
14:30 – 14:45	Break
15:00 – 16:00	Hand-on exercise (2)
16:00 – 16:30	Day 2 Wrap-up


Statistical test
Parametric
Non-Parametric

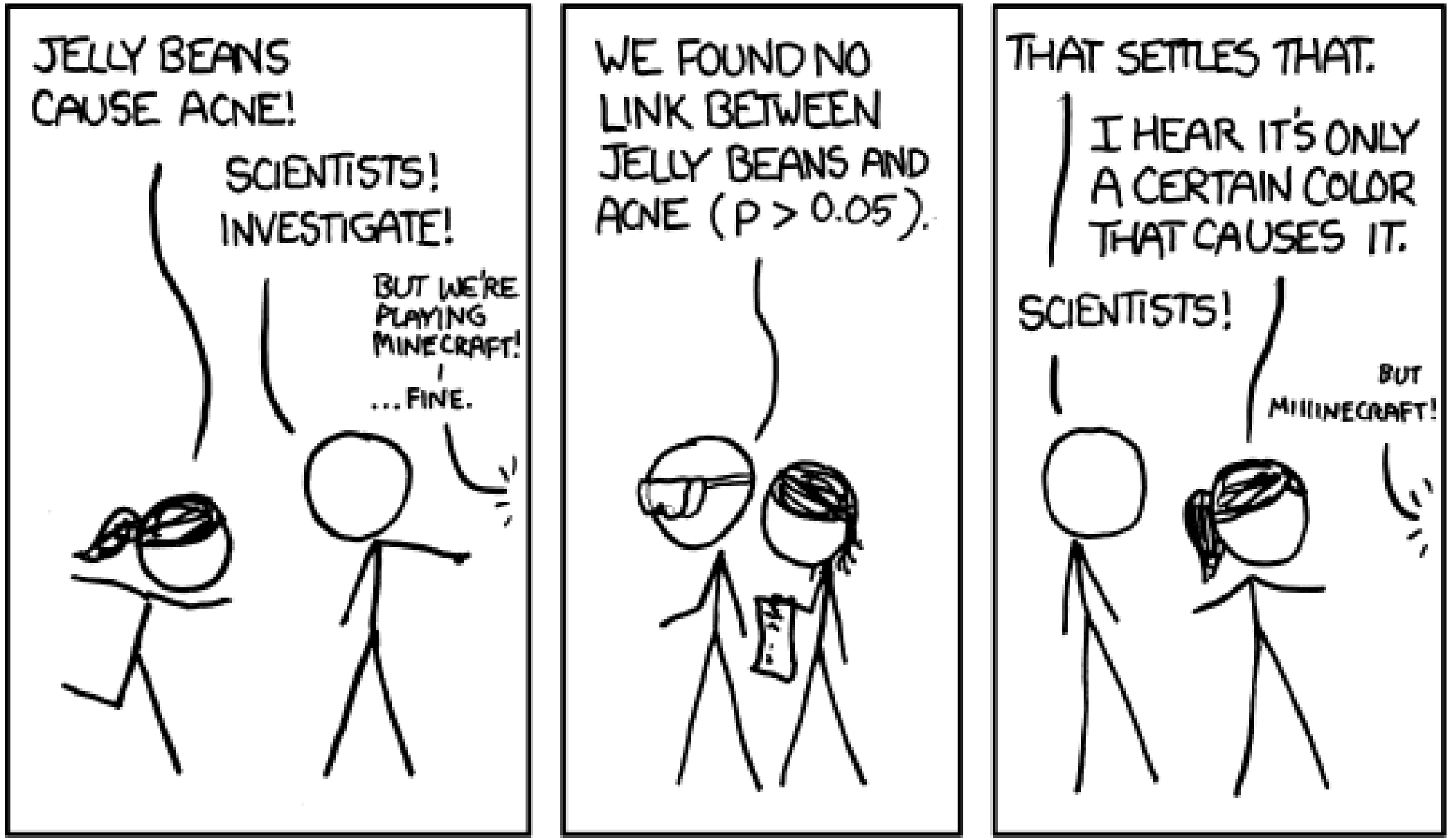
Statistical Significant

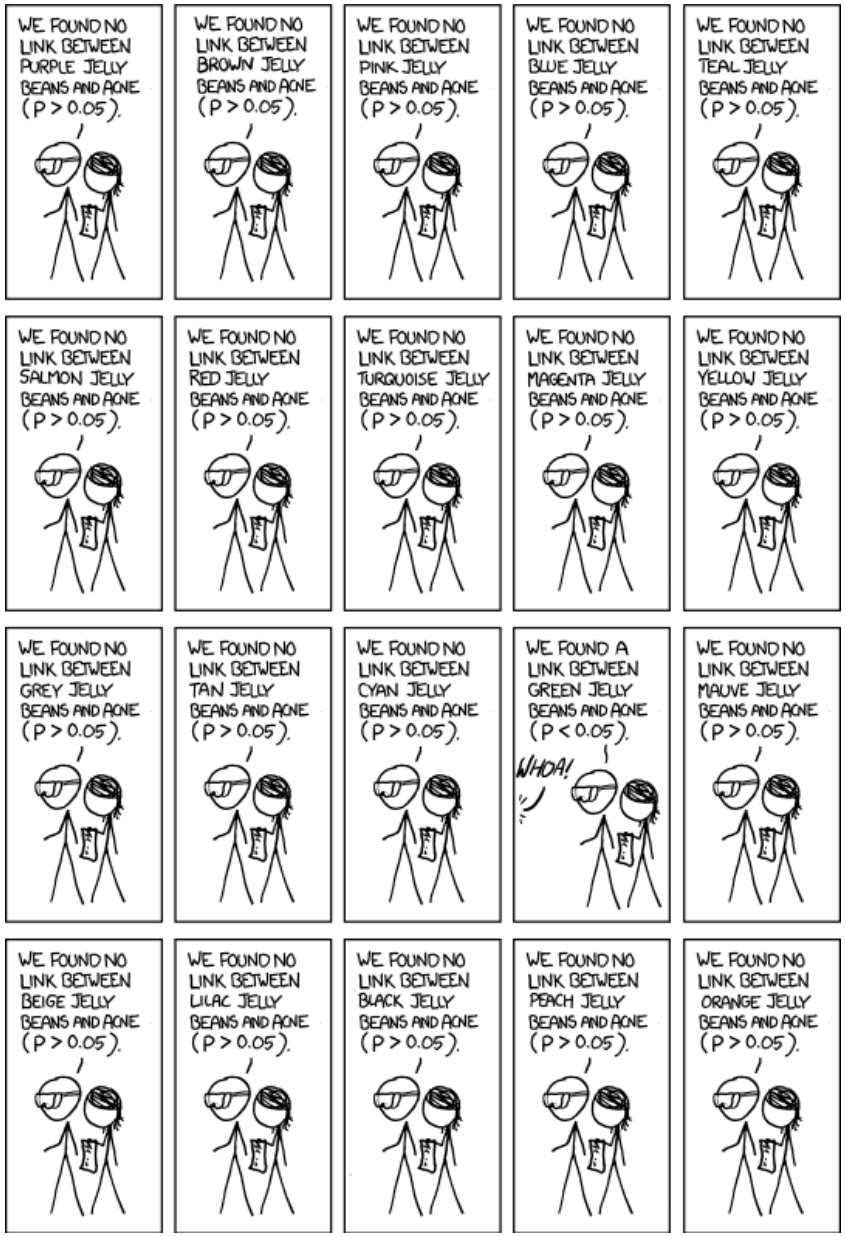


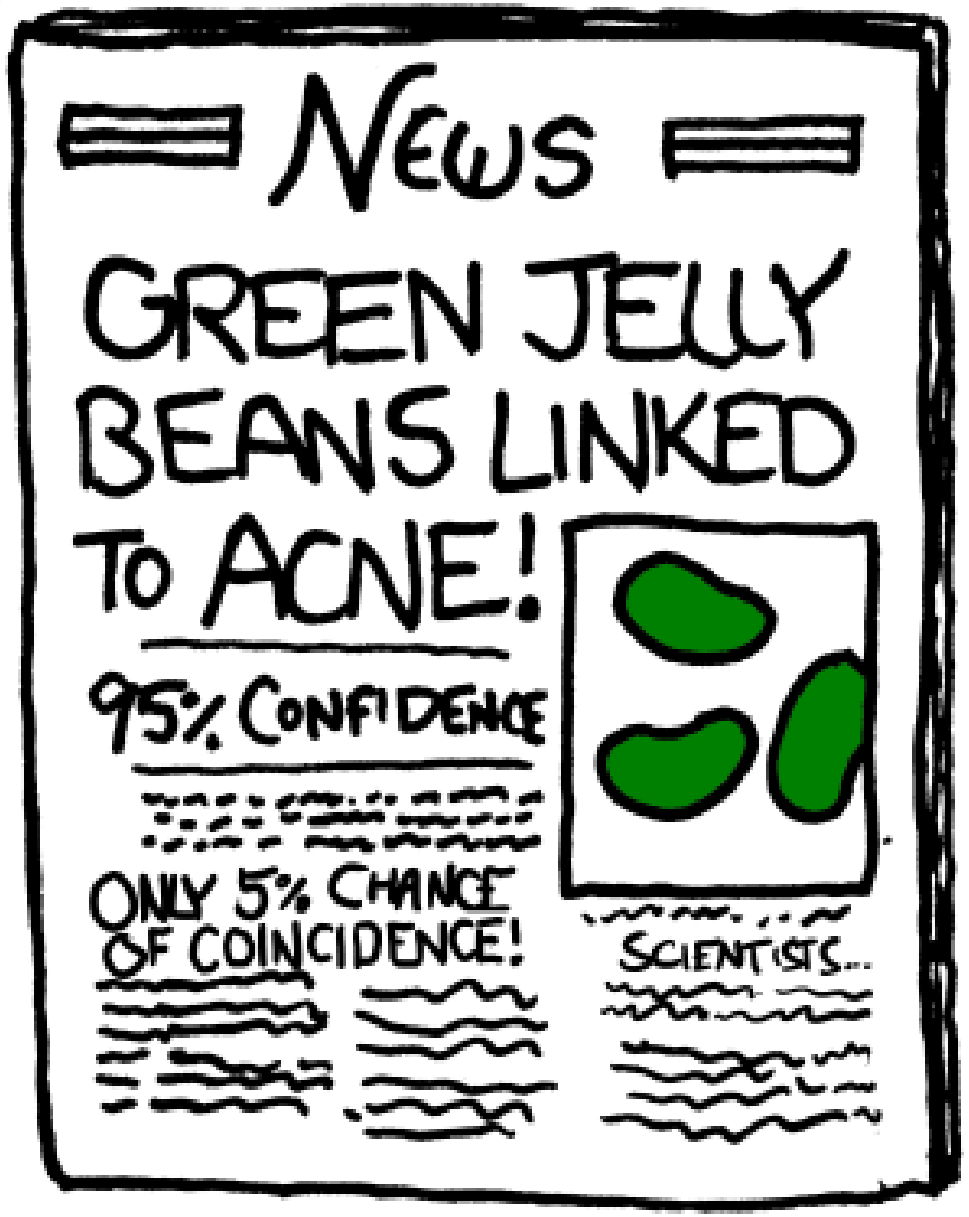
Substantive Significant

Statistical vs. Substantive significant

- Ideally, we want both.
- Statistical significant based on p-value and hypothesis testing.
- Substantive significant is based on our knowledge of the world. What is worth to telling people about.


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 MILLINECRAFT!


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).
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).
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).
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).


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

Risks of and risk factors for COVID-19 disease in people with diabetes: a cohort study of the total population of Scotland

Stuart J McGurnaghan, Amanda Weir, Jen Bishop, Sharon Kennedy, Luke A K Blackbourn, David A McAllister, Sharon Hutchinson, Thomas M Caparrotta, Joseph Mellor, Anita Jeyam, Joseph E O'Reilly, Sarah H Wild, Sara Hatam, Andreas Höhn, Marco Colombo, Chris Robertson, Nazir Lone, Janet Murray, Elaine Butterly, John Petrie, Brian Kennon, Rory McCrimmon, Robert Lindsay, Ewan Pearson, Naveed Sattar, John McKnight, Sam Philip, Andrew Collier, Jim McMenamin, Alison Smith-Palmer, David Goldberg, Paul M McKeigue, Helen M Colhoun, Public Health Scotland COVID-19 Health Protection Study Group, Scottish Diabetes Research Network Epidemiology Group

Findings Of the total Scottish population on March 1, 2020 (n=5 463 300), the population with diabetes was 319 349 (5·8%), 1082 (0·3%) of whom developed fatal or critical care unit-treated COVID-19 by July 31, 2020, of whom 972 (89·8%) were aged 60 years or older. In the population without diabetes, 4081 (0·1%) of 5 143 951 people developed fatal or critical care unit-treated COVID-19. As of July 31, the overall odds ratio (OR) for diabetes, adjusted for age and sex, was 1·395 (95% CI 1·304–1·494; p<0·0001, compared with the risk in those without diabetes. The OR was 2·396 (1·815–3·163; p<0·0001) in type 1 diabetes and 1·369 (1·276–1·468; p<0·0001) in type 2 diabetes. Among people with diabetes, adjusted for age, sex, and diabetes duration and type, those who developed fatal or critical care unit-treated COVID-19 were more likely to be male, live in residential care or a more deprived area, have a COVID-19 risk condition, retinopathy,

What is the statistical significant?

Risks of and risk factors for COVID-19 disease in people with diabetes: a cohort study of the total population of Scotland

Stuart J McGurnaghan, Amanda Weir, Jen Bishop, Sharon Kennedy, Luke A K Blackbourn, David A McAllister, Sharon Hutchinson, Thomas M Caparrotta, Joseph Mellor, Anita Jeyam, Joseph E O'Reilly, Sarah H Wild, Sara Hatam, Andreas Höhn, Marco Colombo, Chris Robertson, Nazir Lone, Janet Murray, Elaine Butterly, John Petrie, Brian Kennon, Rory McCrimmon, Robert Lindsay, Ewan Pearson, Naveed Sattar, John McKnight, Sam Philip, Andrew Collier, Jim McMenamin, Alison Smith-Palmer, David Goldberg, Paul M McKeigue, Helen M Colhoun, Public Health Scotland COVID-19 Health Protection Study Group, Scottish Diabetes Research Network Epidemiology Group

Interpretation Overall risks of fatal or critical care unit-treated COVID-19 were substantially elevated in those with type 1 and type 2 diabetes compared with the background population. The risk of fatal or critical care unit-treated COVID-19, and therefore the need for special protective measures, varies widely among those with diabetes but can be predicted reasonably well using previous clinical history.

What is the substantive significant?

Your hypothesis is …

1. Groups are different from each other

2. Some treatment has an effect on an outcome measure

3. One variable can predict another variable

Null hypothesis (H_0)

There is no difference between a parameter and a specific value, or that there are no differences between two parameters.

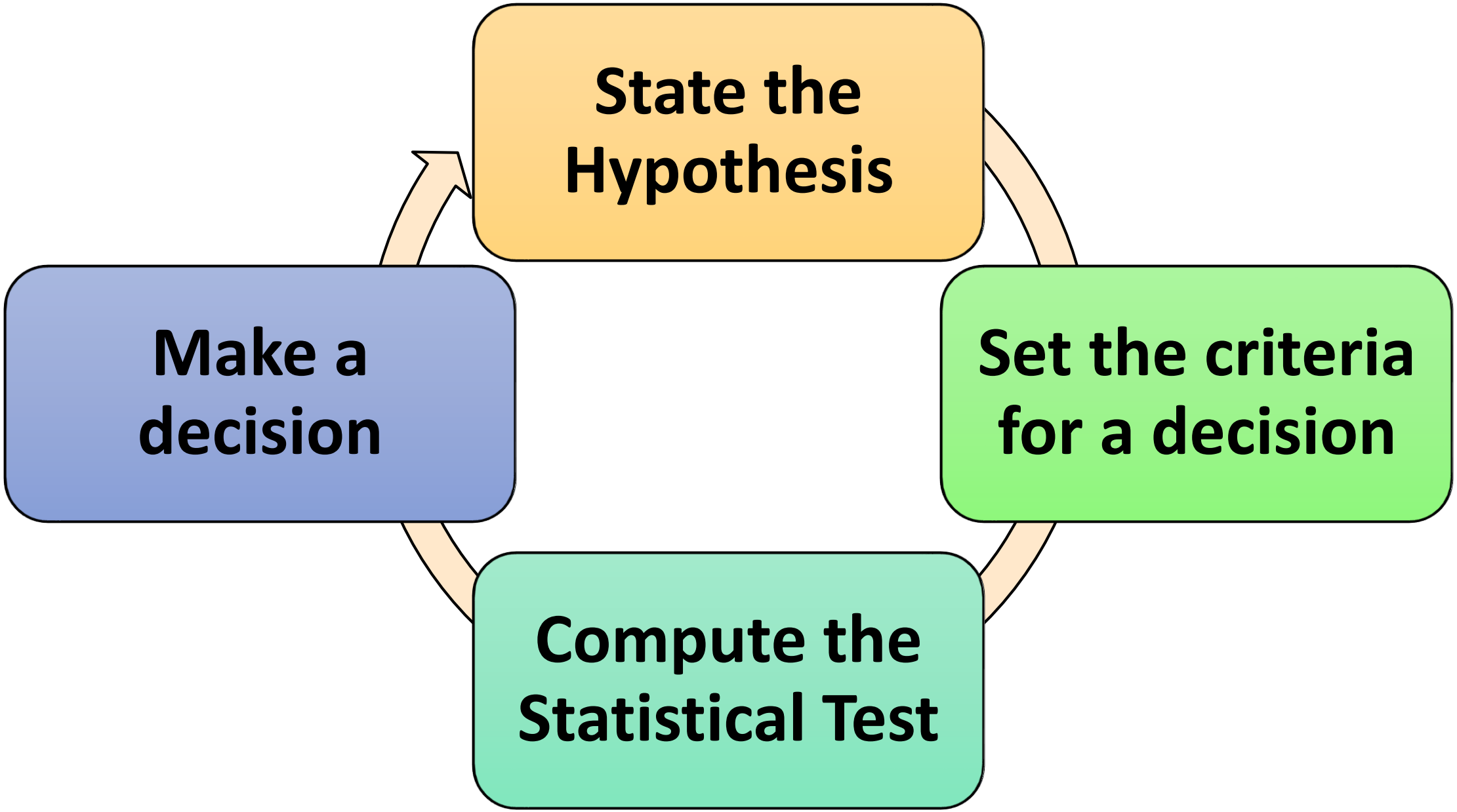
Alternative hypothesis (H_1)

It is a statement that directly contradicts a null hypothesis by stating that that the actual value of a parameter is less than, greater than, or not equal to the value stated in the null hypothesis.

Mathematically

- $H_0: p_1 = p_2$ and $H_1: p_1 \neq p_2$
- $H_0: p_1 = p_2$ and $H_1: p_1 > p_2$ or $H_1: p_1 < p_2$

- $H_0: \mu_1 = \mu_2$ and $H_1: \mu_1 \neq \mu_2$
- $H_0: \mu_1 = \mu_2$ and $H_1: \mu_1 > \mu_2$ or $H_1: \mu_1 < \mu_2$


State the Hypothesis
Set the criteria for a decision
Compute the Statistical Test
Make a decision

		Reality	
	Decision	H_0 True	H_0 False
Outcome	**Do not reject H_0**	No Error $(1-\alpha)$	Type II Error (β)
	Reject H_0	Type I Error (α)	No Error $(1-\beta)$

P-value

$$p - value = \mathbf{Pr}\left(\begin{matrix}\text{Observing a value as or more extreme} \\ \text{than the observed test statistic}\end{matrix} \mid H_0\right)$$

Measuring of evidence

- The smaller p-value, the higher is the "significant" evidence against H_0
- If the p-value is large, we can only say that there is insufficient evidence to reject H_0. We never accept H_0.

If $\boldsymbol{p < \alpha}$, we **reject $\boldsymbol{H_0}$**. If $\boldsymbol{p > \alpha}$, we **do not reject $\boldsymbol{H_0}$**.

One sample test for population proportion

$$H_0: p_1 = p_2 \quad \text{and} \quad H_1: \ p_1 \neq p_2$$

Questions

Suppose a public health practitioner wanted to see if the prevalence of smoking among Thai youths aged 15 to 19 years in 2004 is different from 1999 which is reported to be 6.3%.

He then conducted a large-scale survey and found that 812 out of the 8037 youths he sampled smokes.

Are his estimates of smoking prevalence in 2004 different from the reported prevalence in 1999 (let's assume the true prevalence is 6.3% in 1999)?

What is your null hypothesis?

$$H_0: p = 0.063 \text{ and } H_1: p \neq 0.063$$

```
>binom.test()
>prop.text()
```


Continue on R

One sample test for population mean

- Please load yesterday data “gbsg.csv”

- Suppose normal patients have the progesterone receptor (pgr) greater 10 fmol/L. Can the patients in this study consider to be normal?

What is your null hypothesis?

$H_0: \mu = 10$ and $H_1: \mu > 10$

```
>t.test()

>binom.text(sum(gbsg_data$pgr>10,na.rm=T),length(gbs
  g_data$pgr))
>wilcox.text()
```

`t.test()`

Parametric test

- The data is believed to be drawn from a **normal distribution**, or if the **sample size** is fairly large.

`wilcox.test()`

Non-Parametric test

- No assumption of **normal distribution**, or large **sample size**.
- Based on observations.

Testing the different in population proportions

$$H_0: p_1 = p_2 \quad \text{and} \quad H_1: \ p_1 \neq p_2$$

Or

$$H_0: p_1 - p_2 = 0 \quad \text{and} \quad H_1: \ p_1 - p_2 = 0$$

Parametric Z-test

- If the data comes from a Normal distribution and you know the population standard deviation *a priori*. We use the z-test.
- However in R there is no z-test function.

$$z = \frac{\frac{x_1}{n_1} - \frac{x_2}{n_2}}{\sqrt{\hat{p} \cdot (1 - \hat{p}) \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

```
z.prop = function(x1,x2,n1,n2){
  numerator = (x1/n1) - (x2/n2)
  p.common = (x1+x2) / (n1+n2)
  denominator = sqrt(p.common * (1-p.common) * (1/n1 + 1/n2))
  z.prop.ris = numerator / denominator
  return(z.prop.ris)
}
```

Parametric Z-test

- If $|z| < 1.96$, then the different **is not significant at 5%**
- If $|z| > 1.96$, then the different **is significant at 5%**

Non-parametric $\chi^2 -$ test

- Suppose we cannot assume that the data approximate a Guassian distribution and we don't know the population standard deviation, we test the difference in proportions with a χ^2-test applied to 2×2 contingency table.

```
prop.test(x = c(x1, x2), n = c(n1, n2))
prop.test(x = c(x1, x2, y1, y2))
```

Questions

Thailand's National Statistics Office found that the smoking prevalence among vocational school students was higher at 16.0% compared to secondary school students at 5.1%.

Suppose a public health practitioner wanted to see if the prevalence of smoking among vocational school and secondary school students in Thailand is the **same**.

In his survey, he found that **324 out of 4987 secondary school students smoke** and **488 out of 3050 vocational school students smoke**. He then conducted a large-scale survey and found that 812 out of the 8037 youths he sampled smokes.

What is your null hypothesis?

$$H_0: p_1 - p_2 = 0 \text{ and } H_1: p_1 - p_2 \neq 0$$

```
>prop.test(x = c(324, 488), n = c(4987, 3050))
```

Testing the difference in population means

$$H_0: \mu_1 = \mu_2 \quad \text{and} \quad H_1: \mu_1 \neq \mu_2$$

Or

$$H_0: \mu_1 - \mu_2 = 0 \quad \text{and} \quad H_1: \mu_1 - \mu_2 = 0$$

Parametric `t.test()`

- Under the assumption that both samples are **random, independent**, and come from **normally distributed population** with ***unknown* variances**.

Fisher's F-test for Equality of Two Variances

- Before proceeding with the t-test, it is necessary to evaluate the sample variances of the two groups, using a Fisher's F-test to verify the homoscedasticity (homogeneity of variances).

- In R you can do this with the `var.test()` function.

Fisher's F-test for Equality of Two Variances

- Before proceeding with the t-test, it is necessary to evaluate the sample variances of the two groups, using a Fisher's F-test to verify the homoscedasticity (homogeneity of variances).

$$H_0: \sigma_1^2 = \sigma_2^2 \quad \text{and} \quad H_1: \sigma_1^2 \neq \sigma_2^2$$

```
>var.test(sample1,sample2)
>t.test(sample1,sample2, var.equal=TRUE,
  paired=FALSE)
```

Non-parametric test

- Suppose we cannot assume that samples are taken from populations that follow a Gaussian distribution, we compare the means of the groups with a Mann-Whitney U-test:

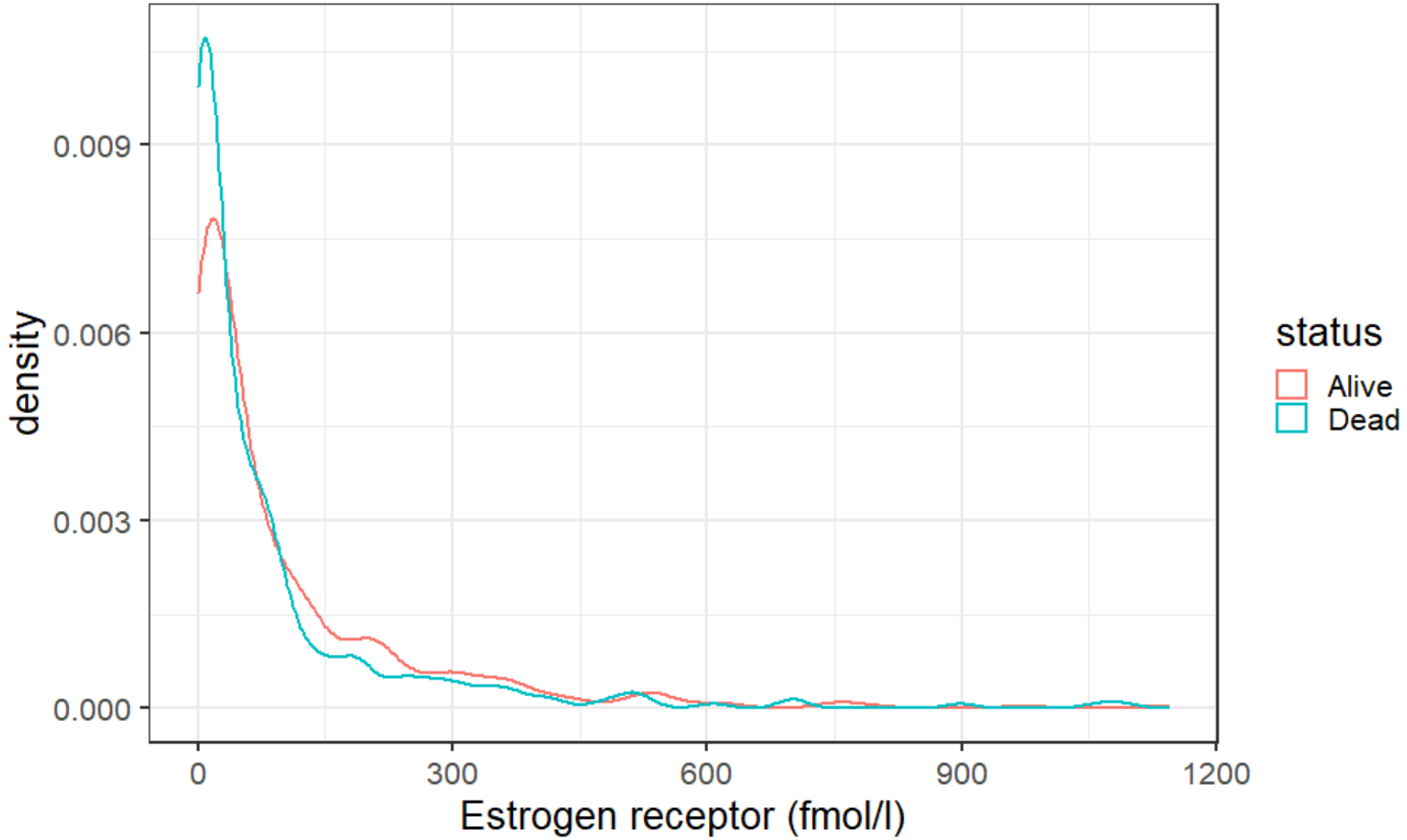
```
>wilcoxon.test(sample1,sample2, paired=FALSE)
```

Question

- Suppose 299 patients from the total of 689 patients was died after hormonal therapy.
- To determine the outcome of hormone receptor level as the main risk factor of the hormonal therapy.

- **Are the estrogen receptor level (er) are the same in the patients who died and survived?**

Plot the estrogen receptor level



Question

- Suppose 299 patients from the total of 689 patients was died after hormonal therapy.
- To determine the outcome of hormone receptor level as the main risk factor of the hormonal therapy.

- **Are the estrogen receptor level (er) are the same in the patients who died and survived?**

What is your null hypothesis?

$H_0: \mu_1 = \mu_2$ and $H_1: \mu_1 \neq \mu_2$

```
>var.test(gbsg_data$er[gbsg_data$status=="Alive"],
 gbsg_data$er[gbsg_data$status=="Dead"])
```

```
># Parametric Test
>t.test(gbsg_data$er[gbsg_data$status=="Alive"],
 gbsg_data$er[gbsg_data$status=="Dead"],
 var.equal = TRUE)
```

Summary

- Hypothesis tests are not a replacement for **estimation** and **confidence intervals**.

- p-values are useful but confidence intervals are more commonly reported
- p-values are a **measure of evidence**
- Confidence intervals are a **measure of effect size**
- Statistical and clinical importance