

# GER1000 Notes

## Legend



This is an exam point!



This is a good to know!

## QR Framework

1. Frame
2. Specify
3. Collect
4. Analyse
5. Communicate

## Design of Studies

*Collect, Analyse, Communicate*

### Problem Faced: Effect of Treatment

We want to see the effect of some treatment. There is thus a need compare the treatment effect using a **treatment group** and a **control group**. This eliminates the **effect of time**. This is called a **controlled experiment**.



**Effect of Time:** Sometimes, it may be tempting to use historical controls. However, there are a lot of confounders.

If control and treatment groups are similar, a difference in response can be attributed to the treatment, else the treatment effect may be confounded with some other factor.

But how do we get similar groups? Often, for controlled experiments, subjects who refuse treatment or agree to treatment do so for a reason. There are thus confounding factors. Thus, only subjects who agree should be enrolled and assigned to control or treatment. And the process requires some assignment that makes these two groups as similar as possible.

This leads us to Randomised Assignment.

### Solution: Randomised Controlled Experiments

Using an impartial procedure based on chance, such as random draws without replacement, we can form two groups from the willing subjects randomly.



**Random:** Strict meaning related to an impartial chance mechanism.

The **law of large numbers** ensures that all confounding factors are almost equally present in both groups.



**Randomised controlled experiments** minimise confounding.

### Problem Again: Awareness of one's group

What if the people know whether they're in the control group or treatment group? Sometimes, just the awareness results in differences in psychology and behaviour, which may affect the results.

The same applies for the doctors administering the treatment. There may thus be biases in the doctors' actions and evaluations, e.g. they may have a higher tendency to diagnose a patient to have been cured should they know that the patient is actually in the treatment group.

This leads us to Double-Blind Experiments.

### Solution: Double-Blind Experiments

In double-blind experiments, both the subjects and the doctors do not know which group each subject is in. This guards against bias.

The subjects would thus be treated with a **placebo**.

Occasionally, this may not be possible. We would thus have to settle for single-blind.


### Problem Returns: Adverse Consequences

What if the "treatment" is something negative, such as smoking? Controlled experiments are no longer possible. The only way is to observe the two

groups.

## Solution: Observational Studies

In an observational study, investigators **do not assign** the subjects to the groups.

 However, there are still treatment and control groups. It is a **self-selecting process**.

In these cases, the "treatment" is also called the **exposure variable**, while the outcome, whether the subject gets "treated", is the **disease or response**.

As we lose out on randomness in the self-selecting process, we cannot eliminate confounding factors.

Furthermore, an observational study **can only establish association**.




## Association

### Introduction: Context of an Observational Study


*Collect, Analyse*

We can represent the results of an observational study using a 2x2 table. Note that it's 2x2 because the exposure and disease are both either present or absent, and do not refer to the actual dimensions of the table.

2x2 Table

Aa.	 Diseased	 Not Diseased	 Row Total
Exposed	38	14962	15000
Not Exposed	44	84956	85000
Column Total	82	99918	100000

There are thus two ways to show association.


 **Disease (X) and Exposure (Y) are associated if**


1.  $\text{Rate}(X|Y) \neq \text{Rate}(X|\text{not } Y)$ , or
2.  $\text{Rate}(Y|X) \neq \text{Rate}(Y|\text{not } X)$

where  $\text{Rate}(X|Y) = ((X \text{ and } Y) \div Y) \times 100\%$ .


Using the table above,  $\text{Rate}(\text{Exposed}|\text{Diseased}) = 38 \div 82 \times 100\% = 46.3\%$ , while  $\text{Rate}(\text{Exposed}|\text{Not Diseased}) = 14962 \div 99918 \times 100\% = 15.0\%$ .

Since the two rates are not equal, there is an association between the exposure and the disease.

 If  $\text{Rate}(X|Y) \neq \text{Rate}(X|\text{not } Y)$ , then  $\text{Rate}(Y|X) \neq \text{Rate}(Y|\text{not } X)$ .


 **Direction of Association**

- If  $\text{Rate}(X|Y) > \text{Rate}(X|\text{not } Y)$ , then X and Y are positively associated.  
If  $\text{Rate}(X|Y) < \text{Rate}(X|\text{not } Y)$ , then X and Y are negatively associated.

 **Combining Rates**

If  $\text{Rate}(A|B) = x$  and  $\text{Rate}(A|C) = y$  with  $x \neq y$ , then overall  $r(A|B+C)$  is between x and y.


### Problem: Confounding

 A **confounder** is a factor associated with both the exposure and the disease.

Confounders obscure the relationship between the exposure and the disease.

 **Association is not causation.**

Since confounders may be involved, association is not causation.

 A confounder **must be different** from the exposure and the disease.

## Solution: Slicing

We can control for the confounder by slicing "along" the confounder. For example, if gender is a confounder, then we can slice some results into two groups based on gender.

This results in smaller groups which are **relatively homogeneous** with respect to the confounding factor.

Other more sophisticated statistical methods include regression.

## Possible Effect of Confounding: Yule-Simpson Paradox

The Yule-Simpson Paradox occurs when relationships between subgroups are reversed when you combine the subgroups.

In other words, if in **most** groups  $\text{Rate}(A|B) > \text{Rate}(A|\text{not } B)$ , but the combined group has  $\text{Rate}(A|B) \leq \text{Rate}(A|\text{not } B)$ , then Yule-Simpson Paradox has occurred.



If the Yule-Simpson Paradox is observed, **then there is confounding**. However, **the reverse may not be true**; not all confounding results in the paradox.

## More on Association: Statistical Relationship

*Analyse, Communicate*

In studies, we're interested to find an average pattern of one variable in terms of another. This would help us to figure out **non-deterministic relationships**.



A **deterministic relationship** allows for precise determination of one variable's true value from another, usually using a formula.

## Analysing Bivariate Data



**Bivariate data** refers to a data set with each data point containing values of two variables.

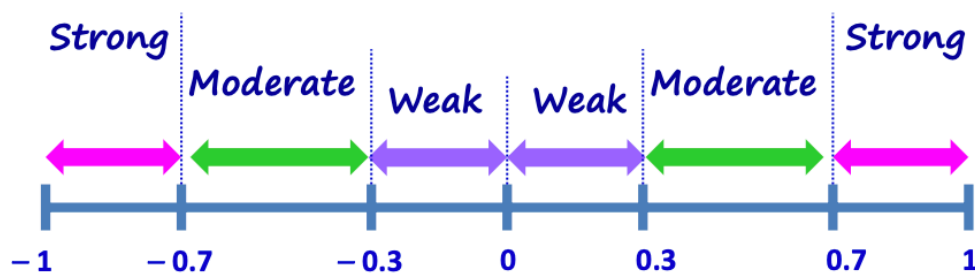
There are a few ways to analyse bivariate data:

1. **Average:** We can find the averages for the two variables. However, this is not of much significance.
2. **Standard Deviation:** This describes the spread or variability of the data around the average.
3. **Scatter Diagram:** We can plot a scatter diagram with the data points.
  1. **45° line:** If the two variables are of the same scale, we can draw a 45° line to see the distribution of points above and below the line. If it's above, then  $x < y$ , while the opposite means  $x > y$ .
  2. **Shape:** We can analyse the shape of the "cloud" formed by the points. If the shape is elongated along the top-right to bottom-left line, then there is positive association between the two variables. If it is elongated along the top-left to bottom-right line, then there is a negative association.
4. **Correlation Coefficient:** Quantifies the relationship. More on this in the next section.
5. **Linear Regression:** Derives a line of best fit to data. More on this in a later section.

## Quantifying Linear Relationships: Correlation Coefficient

The correlation coefficient,  $r$ , is a measure of **linear** association between two variables, where  $-1 \leq r \leq 1$ , with no units. Thus, it shows both direction and strength. Graphically, it refers to how close the data points are to a **straight line form**.

Interpreting the correlation coefficient:



1.  **$r > 0$ :** There is a positive association.
2.  **$r < 0$ :** There is a negative association.
3.  **$r = 0$ :** No **linear** association. There may still be a non-linear association.
4.  **$r = 1$ :** Perfect positive association. All points lie along a single straight line.

5. **r = -1**: Perfect negative association. All points lie along a single straight line.



In the first two cases, the actual relationship **may not be linear**! It can be a curve that also displays some linear association. We need to look at the scatter diagram. High *r* does not mean it's linear.

To compute *r*, we need to do the following for each data point:  $((x - \text{average}_x) \div \text{standard\_deviation}_x)((y - \text{average}_y) \div \text{standard\_deviation}_y)$ . We then need to get the average of this value for all data points.

In other words, if there are 1078 data points,

$$r = \frac{1}{1078} \sum_{i=1}^{1078} \left( \frac{X_i - \bar{X}}{sd_X} \right) \left( \frac{Y_i - \bar{Y}}{sd_Y} \right)$$



**Switching** the two axes/variables **does not change** the correlation coefficient.



**Adding a constant** to all values of a variable **does not change** the correlation coefficient.



**Multiplying a positive** number to all values of a variable **does not change** the correlation coefficient.

### Limitations of Correlation Coefficient

Firstly, **correlation does not mean causation**. This is an extension of the previous association does not mean causation.

Secondly, **outliers** may drastically increase or decrease correlation coefficients. They can be identified using a scatter diagram (in this module). However, we cannot remove them without understanding them, as they may be telling us something important.

Thirdly, correlation only tells us information about linear association, even if  $r > 0$  or  $r < 0$ . The variables may still be **associated in other ways**.

### Correlation Fallacy: Ecological

Often, we have ecological correlation, which is correlation based on aggregated data.



**Ecological fallacy** occurs when we deduce inferences on correlation between individuals based on aggregated data.

If ecological and individual correlations are in the same direction, we will end up **overstating** the strength of the association in individuals, since all variations in individual data are eliminated when aggregated.

They may even be in opposite directions! That's even worse.

### Correlation Fallacy: Atomistic



**Atomistic fallacy** occurs when we generalise the correlation based on individuals towards the aggregate-level correlation.

### Correlation Fallacy: Overgeneralisation



**Overgeneralisation** occurs when we generalise a correlation based on a **subset** of individuals towards a larger set of individuals. No aggregation is occurring here.

### Correlation Phenomena: Attenuation Effect



**Attenuation effect** is what happens when we restrict the range of one variable and cause the correlation coefficient to **decrease in strength**. The scatter plot should be an oval shape for this to occur.



When this occurs, it means that both variables follow a **normal distribution**. The scatter plot should thus be an oval shape.

### Correlation Phenomena: Removal of Data Points

Don't.

## Predicting Values: Linear Regression

If we can find a best fit line to a certain set of data, we can use it to predict values. This line is called the regression line.

$$Y = a + bX$$

The determination of a and b is done using the **least squares method**. Y is the prediction, X is a predictor. We use X to predict Y.



The Y predicted is an average value for a given X.



We **cannot** use it to predict Y using a value of X that is beyond the range in the data set. In other words, **extrapolation is not okay**.



Gradient, b, is not the same as the correlation coefficient, r, generally.

## Sampling

### Introduction: Terminology

**Units:** Elements from which measurements are taken.

**Population:** Collection of all units.

**Parameter:** A numerical fact about a population.

**Sample:** Subset of the population used for study, in order to estimate a population parameter.

**Sampling Unit:** An intermediate unit for sampling. For example, the actual unit may be people, and a sampling unit may be the address of the person, or the doctor this person visits etc. The simplest sampling unit would be the unit itself.



A sampling unit is **not the same** as a population unit.

**Sampling Frame:** A list of sampling units intended to identify all units in the population.

**Census:** A study of the entire population of interest. There is no need for 100% response rate.

### Good Sample: How and Why?

The **purpose** of a sample is so that the results obtained **can be extended** to the population from which the sample was drawn.

For this to be possible, **every unit must have a possibility of being selected**.

The possibility **must be equal**, i.e. selection process is not biased.



As long as **sampling is being done instead of a census**, the limitation of subjects not being representative of the population will hold.

### Good Sampling Frame: How?

A good sampling frame covers **exactly or bigger** than the target population such that every unit has a chance of being selected. However, if it's too big, it may be costly to filter subsequently.

It must also be **up-to-date and complete**.



**Imperfect sampling frames** can result in too many unwanted units, increasing the cost of study. If it excludes desired units, we would need to either redefine the target population, or assess the impact of excluding those units.

## Probability Sampling: Best Type of Sampling Plans



Every unit in the population must have a **known non-zero probability** of being selected into the sample.



There is no need for this probability to be equal!

This ensures that there are no external influences in our selection process.

### Simple Random Sampling (Probability)

Every possible sample of the same size has the same chance of being selected.

## Systematic Sampling (Probability)

Let's say the sample size is N, and the  $K = \text{population size} \div N$ . Then a **random number** from 1 to K is selected, and every K-th number from there is also selected to be part of the sample.



The **starting selection must be random**.



If the sampling units are arranged randomly, we can sometimes treat it as a **simple random sampling plan**.

However, if there are **existing patterns** within the arrangement of the sampling units, we may end up with an undesirable sample.



Systematic sampling is useful when you **don't know the number of sampling units** in the population.

## Stratified Sampling (Probability)

The population is first divided into groups or strata, then we perform **probability sampling** (can be any probability sampling plan) on each of them.

## Multistage Sampling (Probability)

As we may take several stages of selection from the sampling unit to the population unit, we can apply **probability sampling** (can be any) at each stage.

Example would be first doing probability sampling on a list of addresses, then doing probability sampling on the list of people living at that address.

## Cluster Sampling (Probability)

We do probability sampling on the sampling unit level, and just take all population units in that sample unit.

Using the previous example, once we select an address, we would just select everyone living at that address.

## Non-Probability Sampling: GG

When non-probability sampling plans are used, generally the results are not able to be extended from the sample to the population, as some biases are involved.

### Volunteer / Self-selected Sampling (Non-Probability)

Simply ask people to volunteer to participate in the study. It is biased as normally only people **with strong views** would bother to give their answers.

### Convenience / Haphazard Sampling (Non-Probability)

When you select people who are conveniently available, such as your friends, or people you meet on the streets. It is biased as the information you get from respondents who are easily available **to you** is different than that from hard ones to get.

### Judgment Sampling (Non-Probability)

When the interviewer uses their own discretion to choose respondents who they deem as "typical" or "representative". This will most certainly be biased, since it's based on the opinion of somebody.

### Quota Sampling (Non-Probability)

Each interviewer is given quota to fill based on some categories, e.g. X number of males, Y number of females, etc. They are free to interview anyone they like. Having the proportions of these categories in the sample similar to those in the population does not make the extension of the results derived from the sample to the population better.



Quote sampling **does not simply mean** sampling to meet a quota. There must be quota for **multiple groups or categories**, based on certain characteristics.

## Estimation Equation

$$\text{Estimate} = \text{Parameter} + \text{Bias} + \text{Random Error}$$



Large sample size alone **does not mean** no bias. This is because "large" is not with respect to total population. Generally, >1000 respondents is a large sample.

## Bias: Selection

There is a systematic tendency to exclude one kind of person or another from the sample.

Results from imperfect sampling frames and non-probability sampling methods.

## Bias: Non-Response

Systematic tendency from subjects who do not respond to the study. This is likely caused by differences between the respondents and non-respondents. If the non-response rate is high, then non-response bias is significant.

### Bias: Others

- Phrasing of question
- Tone of interviewer
- Subjects tend to understate undesirable social habits



The tendency for people respond inaccurately is called **response bias**.

### Random Error

Studies of a certain population parameter would likely have estimates that fluctuate around the actual population parameter. This fluctuation is **random error**, due to the sample being taken **randomly**.

### Reducing Random Error

Taking larger sample sizes reduces the random error! If this sample size reaches the actual population size, then random error basically equals zero.

### Confidence Intervals

Confidence interval is the range of values that we are reasonably certain our parameter lies in.

It is also a way to report results while accounting for random error.

We can thus use 95% CI:  $X \pm Y$  to express we are 95% confident that the range from  $X - Y$  to  $X + Y$  contains the population parameter.



95% CI means that if we repeat the experiment with **different samples but of the same size**, almost 95% will contain the true population parameter.

With reference to random error, since increasing the sample size decreases random error, it also decreases the size of the confidence interval, i.e. more certain of where the parameter is.



**Larger sample size** results in a smaller confidence interval.

## More on Observational Studies

### Risk: What is it?

Risk is the **rate** of an **uncertain condition**.

We can use a **2x2 contingency table** for calculation purposes.

#### 2x2 Contingency Table

<u>Aa</u> .	Diseased	Not Diseased	Row Total
<u>Exposed</u>	38	14962	15000
<u>Not Exposed</u>	44	84956	85000
<u>Column Total</u>	82	99918	100000

Thus, the **Risk(Disease) =  $82 \div 100000 = 0.00082$** .



If there's a multi-level contingency table, e.g. 3x4, then to calculate any risks ratio, odds ratio etc., we need to choose the right values to form a 2x2 table. Generally, that'd mean we need to choose a baseline disease group and a baseline exposure group.

### Risk Ratio: Relative Risk

Risk ratio = Risk(Diseased|Exposed)  $\div$  Risk(Diseased|Not Exposed) =  $(38 \div 15000) \div (44 \div 85000) = 4.8939$ .

This is also known as Relative Risk.



**RR > 1:** The first group has greater risk.

**RR < 1:** The second group has greater risk.

**RR = 1:** There is no association between the disease and the exposure.

### Odds: What is it?

Unlike risks, which is calculated by the rate of a disease, odds is calculated by Odds(Disease) = Risk(Disease)  $\div$  (1 - Risk(Disease)).

In other words, it's  $\text{Number}(\text{Disease}) \div \text{Number}(\text{Not Diseased})$ .

Similarly,  $\text{Odds}(D|E) = \text{Risk}(D|E) \div (1 - \text{Risk}(D|E))$ .

Using the contingency table above, **Odds(Disease) =  $82 \div 99918 = 0.00082067$** .

## Odds Ratio

Odds Ratio =  $\text{Odds}(\text{Diseased}|\text{Exposed}) \div \text{Odds}(\text{Diseased}|\text{Not Exposed})$ .

With just the odds ratio, one is unable to calculate the risk ratio.



**OR = 1:** No difference in disease risk between two groups,  $RR = 1$

**OR > 1:** Higher risk in first group,  $RR > 1$

**OR < 1:** Lower risk in first group,  $RR < 1$



You can use a cross-product-ratio to calculate the estimated odds ratio. Let the disease be the first column, and the exposure be the first row. Then  $OR = (\text{top left} \times \text{bottom right}) \div (\text{bottom left} \times \text{top right})$ .

## Summary Slide: Risks and Odds

# More on observational studies

A simple example of a cohort study is given below, with risks and odds calculations.

	Disease (D)	No Disease	Row Sum
Males (M)	100	200	300
Females (F)	200	600	800
Column Sum	300	800	1100

$$\text{risk}(D|M) = \frac{100}{300} \approx 0.33$$

$$\text{Odds(disease) among males} = \frac{100}{200} = 0.50$$

$$\text{risk}(D|F) = \frac{200}{800} = 0.25$$

$$\text{Odds(disease) among females} = \frac{200}{600} \approx 0.33$$

$$\text{Population RR} = \frac{\text{risk}(D|M)}{\text{risk}(D|F)} \approx \frac{0.33}{0.25} \approx 1.33$$

→ Males have higher odds of getting D

→ Risk of males getting D is abt 33% higher

$$\text{OR(disease) between males \& females} = \frac{0.5}{0.33} \text{ (or } \frac{100 \times 600}{200 \times 200}) = 1.5$$

(here I assumed that females is the baseline) (again I assumed that females is the baseline)

## Type of Study: Cohort Study

A cohort study is one where samples are taken from the exposure group. In other words, with reference to the contingency table, we **fix the row totals** and sample accordingly.

After enrolling the subjects, they are then monitored to find out how many in each exposure group have the disease. This is **future-looking**.

It is, however, expensive.



With cohort studies, you **can use the sample risk ratio to estimate the population risk ratio**.



With cohort studies, you **can use the sample odds ratio to estimate the population odds ratio**.



With cohort studies, **given the odds (not odds ratio), you can calculate the risks**, and vice versa.

## Type of Study: Case Control Study

A case control study is one where samples are taken from the disease groups. In other words, with reference to the contingency table, we **fix the column totals** and sample accordingly.



Case control studies **are not controlled studies**.


Since we are already sampling people who are either diseased or not, this is generally done **at the end of a time period**. Thus, it tends to be **backward-looking**, i.e. it looks at background.


It is less expensive since less time is needed.



It is also **good for rare diseases**, since we cannot expect a cohort study to show significant changes in the number diseased.

 With case control studies, you **can use the sample odds ratio to estimate the population odds ratio**.

 With case control studies, you **cannot use** the sample risk ratio to estimate the population risk ratio (usually).

 With case control studies, you **cannot calculate odds** and also **cannot calculate risk**.

## Summary Slide: Studies

# More on observational studies

## Cohort vs Case Control studies

(OPTIONAL slide to explain why can estimate RR and OR using Cohort study, but can usually only estimate OR using Case Control study)

### Cohort Study

	Disease	No Disease	Row Sum
Males (M)	100 <sub>a</sub>	200 <sub>a</sub>	300 <sub>a</sub>
Females (F)	200 <sub>b</sub>	600 <sub>b</sub>	800 <sub>b</sub>

Let a and b be the proportions of males and females sampled

$$\text{Est. risk}(D|M) = \frac{100a}{300a} \approx 0.33 \text{ (pop. risk} = 0.33)$$

$$\text{Est. risk}(D|F) = \frac{200b}{800b} = 0.25 \text{ (pop. risk} = 0.25)$$

$$\Rightarrow \text{Est. RR} \approx \frac{0.33}{0.25} \approx 1.33 \text{ (pop. RR} = 1.33)$$

$$\text{Est. odds}_{\text{males}} = \frac{100a}{200a} = 0.50 \text{ (pop. odds} = 0.50)$$

$$\text{Est. odds}_{\text{females}} = \frac{200b}{600b} \approx 0.33 \text{ (pop. odds} = 0.33)$$

$$\Rightarrow \text{Est. OR} \approx 1.50 \text{ (pop. OR} = 1.50)$$

### Case Control

	Disease (D)	No Disease	Row Sum
Males (M)	100 <sub>c</sub>	200 <sub>d</sub>	100 <sub>c</sub> + 200 <sub>d</sub>
Females (F)	200 <sub>c</sub>	600 <sub>d</sub>	200 <sub>c</sub> + 600 <sub>d</sub>

Let 300c and 800d be the number in D and no D groups sampled (usually no. of people in D and no D groups not known, hence  $c \neq d$ )

$$\text{Est. risk}(D|M) = \frac{100c}{100c+200d} (\neq \text{pop. risk unless } c = d)$$

$$\text{Est. risk}(D|F) = \frac{200c}{200c+600d} (\neq \text{pop. risk unless } c = d)$$

$$\Rightarrow \text{Est. RR} \neq \text{pop. RR unless } c = d$$

$$\text{Est. odds}_{\text{males}} = \frac{100c}{200d} (\neq \text{pop. odds unless } c = d)$$

$$\text{Est. odds}_{\text{females}} = \frac{200c}{600d} (\neq \text{pop. odds unless } c = d)$$

$$\text{BUT Est. OR} = \frac{100c \times 600d}{200c \times 200d} = 1.5!!!!$$

## Uncertainty

### Introduction: Terminology

**Random circumstance** is one in which an outcome is uncertain, and not determined until we observe it.

**Probability** is a measure of how likely something will happen. It is a value between 0 and 1. An everyday word for it may be **chance**.

**Relative frequency** is one interpretation of probability, and can be quantified exactly. It is based on repeated observation of outcomes.


**Personal probability** is another interpretation of probability, and cannot be quantified exactly. It is based on our own personal belief.

### Relative Frequency as Probability

There are two ways to this:

1. Make an assumption about the physical world to define relative frequency, e.g. a coin lands on heads 50% of the time, and tails the other 50% of the time.
2. Observe relative frequencies over many repetitions to estimate probability, e.g. there's a 70% chance of rain based on past scenarios.

### Complement Rule

  $P(\text{Event}) = \text{Probability of an event}$

$$P(A^C) = 1 - P(A)$$

### Addition Rule

For mutually exclusive events,

$$P(A \cup B) = P(A) + P(B)$$

For events that are not mutually exclusive,

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

### Multiplication Rule

For independent events,

$$P(A \cap B) = P(A) \cdot P(B)$$

For events that are not independent,

$$P(A \cap B) = P(A) \cdot P(B|A)$$

## Conditional Probability

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$



$P(A|B)$  generally does not equal  $P(B|A)$ .



Make sure to check that  $P(A \cap B)$  has accounted for **both  $P(A)$  and  $P(B)$** . It's often easy to miss out one of them or take  $P(B|A)$  as  $P(A \cap B)$ .

By the consistency rule,

$$\text{if } P(A|B) > P(A), \text{ then } P(B|A) > P(B)$$

## Relation to Odds

$$Odds(E) = \frac{P(E)}{P(E^C)}$$

## Average Values

$$V_{\text{avg}} = V(A) \cdot P(A) + V(B) \cdot P(B) + \dots$$

## Hypothesis Testing

Occasionally, we observe a phenomena and ask if it's by chance. We thus come up with two hypotheses.



**Null hypothesis** states that the observation is purely due to chance.



**Alternate hypothesis** is the negation of the null hypothesis.

To see if an observation was simply due to chance, we calculate the probability of the observation + the probability of all equally extreme observations + the probability of more extreme observations, **assuming Null is true**.

This total probability is the p-value.

If the p-value is low enough, usually  $< 0.05$ , then we can reject the null hypothesis, else **we don't know if it's due to chance or not**.

$$\text{P-value} = P(\text{observation} \mid \text{null is true}) + P(\text{equally extreme outcomes} \mid \text{null is true}) + P(\text{outcomes that are more extreme} \mid \text{null is true})$$



**Not rejecting** the null hypothesis **does not mean** that the null hypothesis is true.



We will never attempt to reject the alternate hypothesis since p-value cannot determine anything about the alternative hypothesis.

## Testing Rare Events or Diseases

All the following are in relation to tests used to test for rare diseases.

**Base rate:**  $P(\text{Disease})$

**Sensitivity:**  $P(\text{Positive} \mid \text{Disease})$

**Specificity:**  $P(\text{Negative} \mid \text{No Disease})$

## Miscellaneous

### Regression Fallacy

Ascribing a cause where none exists in situations where natural fluctuations exist while failing to account for these natural fluctuations.