# Introduction to Biostatistics

Principles of Surgery

Maxime Turgeon

March 12, 2019

# Topics to Cover

- Introduction to Statistical Inference
  - Hypothesis Testing
- Experimental vs. Observational Studies
  - Confounding
- Diagnostic Testing
- Survival Analysis

# Before we start

The slides can be found on my website: maxturgeon.ca/talks

Things to keep in mind:

- SPSS (or PSPP) and Stata; **not** Excel
- Study design can make or break a study
- Look at your data! (Tables and graphs)
- Consult a statistician:
  - Departmental research assistant
  - Clinical Research Support Unit
  - Hire a graduate student from Community Health and Epidemiology!

# Introduction to Statistical Inference

# Descriptive vs. Inferential Statistics

Statistics can be broadly broken down into two categories:

- *Descriptive statistics*
  - Describe the properties of the observed dataset.
- *Inferential statistics*
  - Infer properties of a general population from properties of the observed dataset.

# An example

- Suppose we collect the age of all the residents in the classroom.
- We can describe the age distribution using the **mean** (28.5 years) and the **standard deviation** (0.5 years).

At this point, we have only performed descriptive statistics.

# From Description to Inference

- Now suppose we want to make inference about a larger population, of which this classroom is a sample.
- This larger population could be:
  - all surgical residents at USask;
  - all PGY1 and PGY2 residents at USask;
  - all Canadian residents taking Surgical Foundations (or its equivalent).
- Note that this classroom is a sample of *all* three populations.

- But the average age may not be representative of a particular population:
  - Other surgical residents who have completed Surgical Foundations are probably *older*.
  - PGY1 and PGY2 residents in other specialties are probably the same age as surgical residents.
  - Other Canadian surgical residents taking Surgical Foundations are potentially *younger*: students enter medical school at an earlier age in Quebec than the rest of Canada.

In other words, to go from sample to population, we need to understand how the sample was generated, and how it relates to the population of interest. Failure to account for this typically leads to **biases**.

# Important concepts

- An **estimand** is a population-level quantity of interest
  - E.g. the average age among Canadian women, the 5-year survival probability after a breast cancer diagnosis.
- An **estimator** is a function that takes as input a *dataset* and outputs a *summary statistic*
  - E.g. the function that computes the mean or the risk ratio from a sample.
- An **estimate** is the value taken by an estimator for a given dataset.

In statistical inference, we observe the *estimate*, and we use it to make inference about the corresponding *estimand*. The mathematical properties of the *estimator* is what allows us to construct confidence intervals and compute p-values.

# Confidence Interval

- A **confidence interval** is a range of "reasonable" values for the estimand of interest.
- It is usually constructed around the estimate that we obtained.
  - More specifically, a 95% confidence interval is such that, for a given dataset, the probability that the confidence interval constructed contains the true value of the estimand is 95%.

# P-value

- It is defined as a conditional probability under the null hypothesis
  - E.g. the hypothetical scenario of no treatment effect.
- The **p-value** is defined as the probability, assuming the null hypothesis holds, that we could obtained an estimate as "extreme" as the estimate we computed on our dataset.

# Common misconceptions

The p-value is often misinterpreted, even by quantitatively sophisticated audiences.

Common misconceptions of the p-value include:

- That it is a probability under chance; **it is a probability under a very specific scenario**.
- That it is the probability of the null hypothesis being true; **it is not**.
- That it is the probability of the *alternative* hypothesis being false; **it is not**.

# PSA

- The confidence interval should be prefered over the p-value.
  - The p-value should **only** be reported alongside a confidence interval.
- The confidence interval always carries more information than the p-value.
  - Its width provides information about the precision of the study.
  - The range of values can tell us about the clinical relevance of the finding.
- **Never** use a p-value to make decision, e.g. implementing a new procedure, removing a variable from a prediction model.
  - You should combine results from multiple studies (i.e. meta-analysis).
  - You should also think about other factors, e.g. clinical relevance, subject-matter knowledge, cost-benefit analysis.

# An example: Hanley *et al*, Lancet (2019)

- RCT on the efficacy and safety of minimally invasive surgery with thrombolysis in intracerebral haemorrhage evacuation (MISTIE).
- 506 patients were randomised to either MISTIE or standard care.
- An modified Rankin Score (mRS) between 0 and 3 is a *good* functional outcome.

| Treatment | mRS score 0-3 | mRS score 4-6 | Proportion |
|---|---|---|---|
| MISTIE | 110 | 139 | 44.18% |
| Standard Care | 100 | 140 | 41.67% |

- The proportion $p_T$ is slightly higher in the treatment group than the proportion $p_C$ in the control group.
- We would like to know whether this difference is likely to be present in the overall population or whether it could be just a fluke of our particular dataset.
- There are three main ways to summarise the results, which leads to three different inference strategies:
  1. **Difference in absolute risks**: $\Delta = p_T - p_C$.
  2. **Risk Ratio**: $RR = p_T / p_C$.
  3. **Odds Ratio**: $OR = \frac{p_T/(1-p_T)}{p_C/(1-p_C)}$.
- Note that the RR and the OR are measures of *relative risk*.

We can compute confidence intervals and p-values for all three test statistics:

| Statistic | Estimate | Confidence Interval | P-value |
|---|---|---|---|
| Risk difference | 2.51% | (-6.7%, 12%) | 0.64 |
| Risk Ratio | 1.06 | (0.86, 1.3) | 0.58 |
| Odds Ratio | 1.11 | (0.77, 1.59) | 0.58 |

# Other test statistics

- In the example above, we had a binary exposure (MISTIE vs. SC) and a binary outcome (+ive vs. -ive functional outcome).
  - In this setting, we contrasted two proportions, using three different metrics.
- When the type of these variables is different, we can use other tests:

| Exposure | Outcome | Test | Non-parametric analogue |
|---|---|---|---|
| Binary | Continuous | $t$-test | Mann–Whitney $U$ test |
| Paired | Continuous | Paired $t$-test | Wilcoxon signed-rank test |
| Categorical | Continuous | ANOVA | Kruskal–Wallis test |
| Categorical | Categorical | Chi-squared test | N/A |

# Some comments i

- All these tests can be performed using standard statistical software (e.g. SPSS, Stata).
- The difference between a *t*-test and a *paired* *t*-test is that there is a natural pairing between the observations.
  - E.g. Pre- and Post-treatment measures on the same patient.
- All three tests with a continuous outcome assume that it follows a normal distribution.
  - When this assumption fails, we can use the non-parametric equivalent.

# Some comments ii

- The non-parametric tests only return a p-value, no confidence interval.
  - For this reason, they should only be used as a last resource. Resampling techniques (e.g. bootstrap) are more advanced techniques that can be used to compute CIs when the distribution is not normal.
- I only present the one-way Analysis of Variance (ANOVA). There is a whole catalogue of variants of the ANOVA (e.g. more than one independent variable, repeated measurements).
  - However, I recommend using regression techniques instead as they are more informative and more flexible.
- The chi-squared test does not make any assumption about the distribution of the outcome or the exposure. Therefore, it is its own non-parametric analogue.

# Experimental vs. Observational Studies

# RCTs as the Gold Standard

- Randomised Controlled Trials are often hailed as the gold standard of clinical epidemiology. This is due to the fact that, when randomisation leads to good balance, *every patient fully complies with the protocol*, and there are no competing events, the association between the treatment and the outcome will be unconfounded.
- However, it is not always possible to study clinically relevant associations through RCTs (typically for ethical reasons), and for this reason, we often have to resort to observational studies:
  - Cohort studies;
  - Case-Control studies.

# Cohort studies

- A **cohort study** is a study that follows individuals over time, recording their exposure status and the occurence rate of outcomes.
  - The following of individuals over time can actually be done retrospectively.
  - A study that looks at a particular point in time is called a *cross-sectional study*.
  - By following people over time, we can reduce bias from *reverse causation*: by observing the sequence of events, we know whether the exposure or the outcome occurs first. This may not always be possible with cross-sectional studies.

# Case-Control studies

- When prevalence is low, cohort studies may require a large sample size to observe "enough" cases.
- A **case-control study** is an observational study where individuals with the outcome of interest are sampled, along with an appropriate control group.
- Since we are oversampling the cases, we cannot estimate prevalence directly, and similarly we cannot estimate absolute risks.

# Confounding

- A **confounder** is a variable that influences (i.e. causes) the outcome variable and that systematically differs between the exposure groups.
- This leads to an apparent (or spurious) relationship between the outcome and exposure variables. Failure to account for all confounders can lead to biased and erroneous inference.
  - E.g. Selection bias, confounding by indication, attrition bias.
- Common confounders in epidemiology:
  - Age
  - Sex
  - Socio-economic status

- Note that RCTs are **not** immune to confounding.
- The most common reasons:
  - Imbalance in small studies
  - Non-compliance
  - Loss to follow-up
- On the other hand, confounding is more of an issue with observational studies.

# Adjusting for Confounding

- There are three main strategies to adjust the inference for confounding:
  1. **Randomisation** (e.g. RCTs).
  2. **Stratification**.
  3. **Weighting**.
- By stratification, we mean computing an estimate for each stratum (e.g. for each age group, for each sex) and combining them into an overall estimate.
  - *Regression* is also a form a stratification and can therefore be used to adjust for confounders.
- Weighting is a technique coming from the field of surveys where we up-weight or down-weight the observations to make the sample look like the general population.
  - The most common weighting method in epidemiology is the *inverse probability of treatment weighting* (IPTW) and its variants.

# An example: Adeyeye *et al*, AJE (2018)

- The study looked at the association between mode of delivery (cesarean vs. vaginal delivery) and the risk of wheezing in early childhood.
- The sample was selected from the Upstate KIDS study, which was established to study the relationship between infertility treatment and child development.
  - For all NY State but excluding NYC, all births following infertility treatments were enrolled along with 3 control births.
  - All multiple births were also enrolled.

# Potential confounders

What are the potential confounders?

- Pregnancy complications;
- Maternal atopy;
- Gestational age;
- Birth weight;
- Smoking during pregnancy.

Other forms of bias include:

- Selection bias, coming from the sampling design;
- Loss to follow-up;
- Outcome misclassification.

# Diagnostic Tests

# Characteristics of a good diagnostic test

- (Relatively) Inexpensive
- (Relatively) Easy to administer
  - Minimal discomfort
- **Sensitive**: correctly identifies *true disease cases*.
- **Specific**: correctly identifies *true non-disease cases*.

# A few formulas

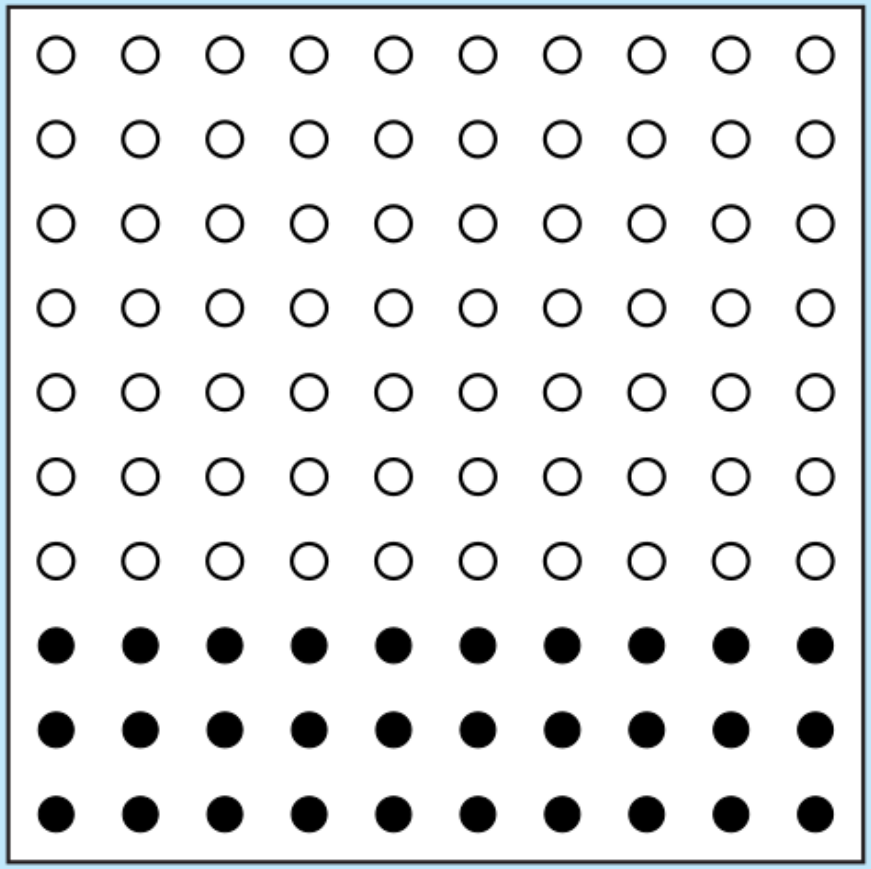
- We assume that patients are given a test to assess whether or not they have a given condition (or disease).

| | Disease | No Disease |
|---|---|---|
| +ive Test | TP | FP |
| -ive Test | FN | TN |

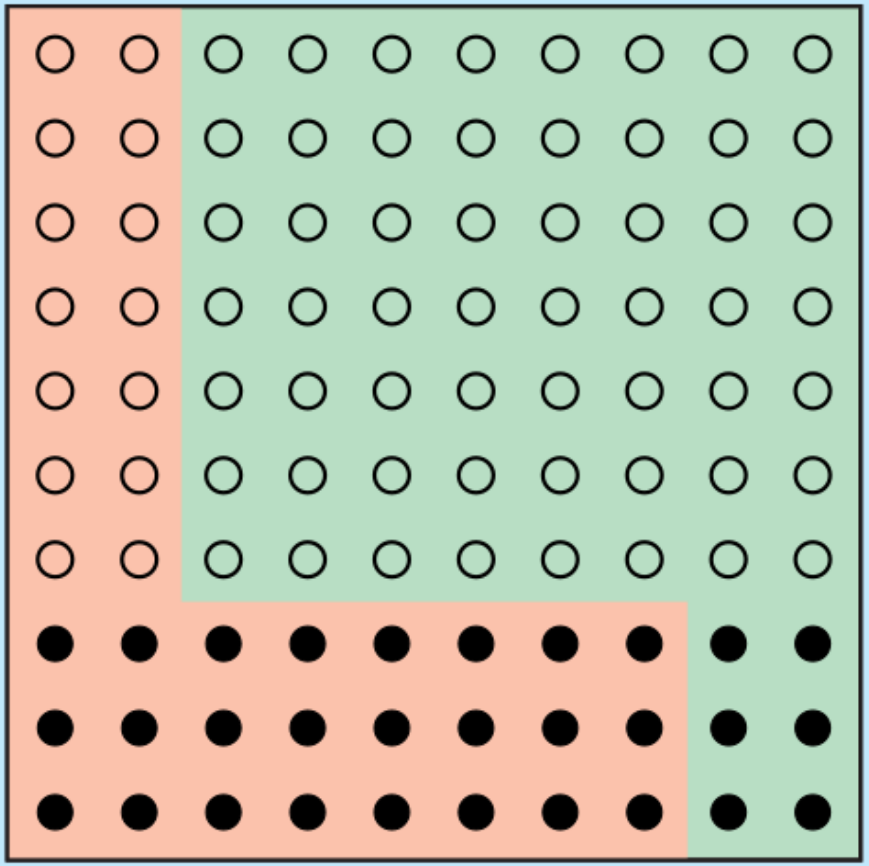
- Sensitivity: $\frac{TP}{TP+FN}$.
- Specificity: $\frac{TN}{TN+FP}$.
- Positive Predictive Value: $\frac{TP}{TP+FP}$.
- Odds Ratio: $\frac{TP/FN}{FP/TN}$.

- In words, this gives:
  - **Sensitivity**: Probability of testing positive, given that the patient has the disease.
  - **Specificity**: Probability of testing negative, given that the patient does not have the disease.
  - **Positive Predictive Value**: Probability of having the disease, given that the patient tested positive.
  - **Diagnostic Odds Ratio**: Relative change in odds of testing positive when patient has disease compared to when they do not.
- Clearly, the most clinically important quantity is the PPV. However, the PPV is influenced by the *prevalence*, i.e. for fixed sensitivity and specificity, the PPV increases the more risk factors the patient has (and **vice-versa**).

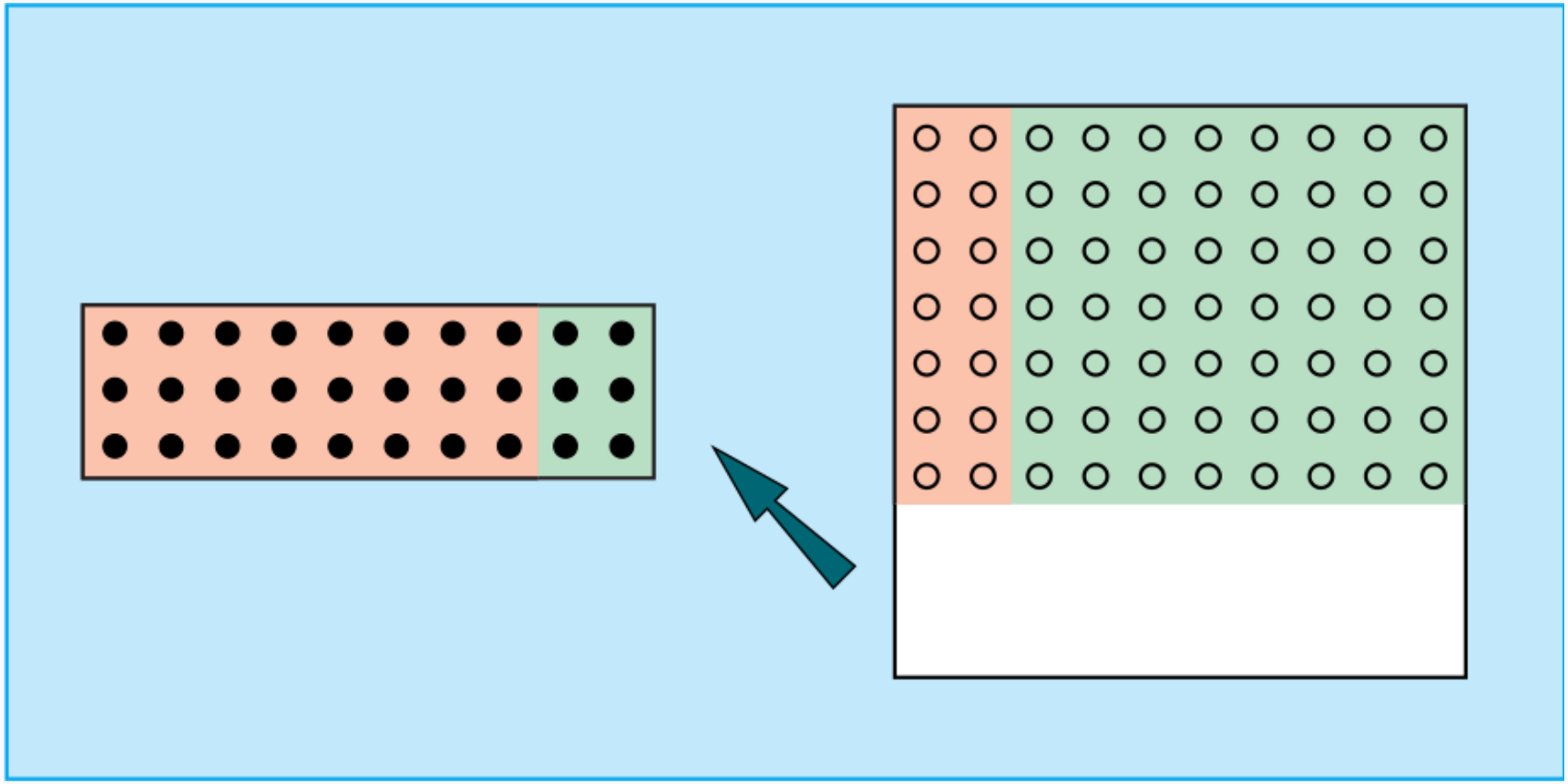
# Hypothetical Population



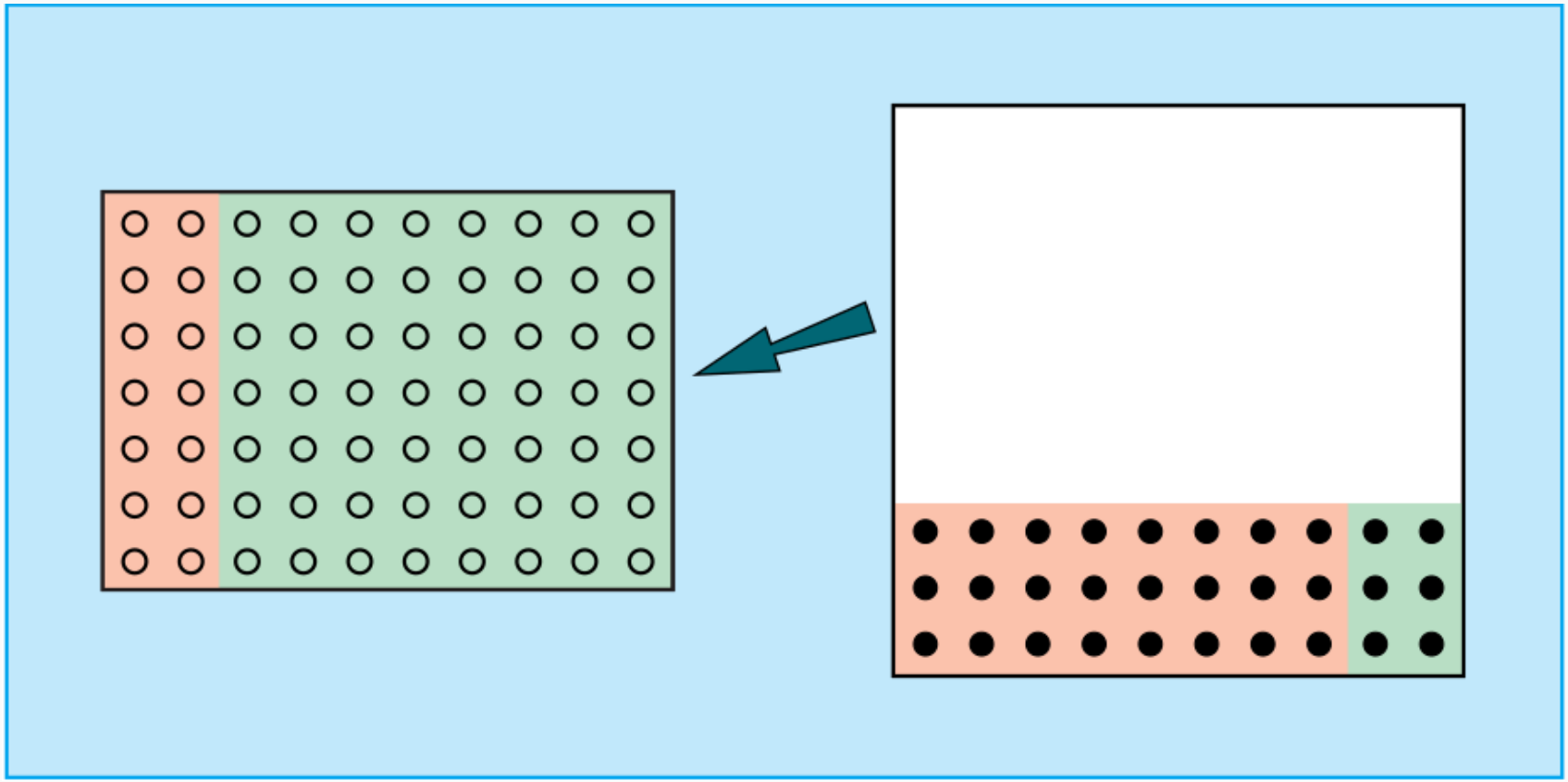
# Hypothetical Test



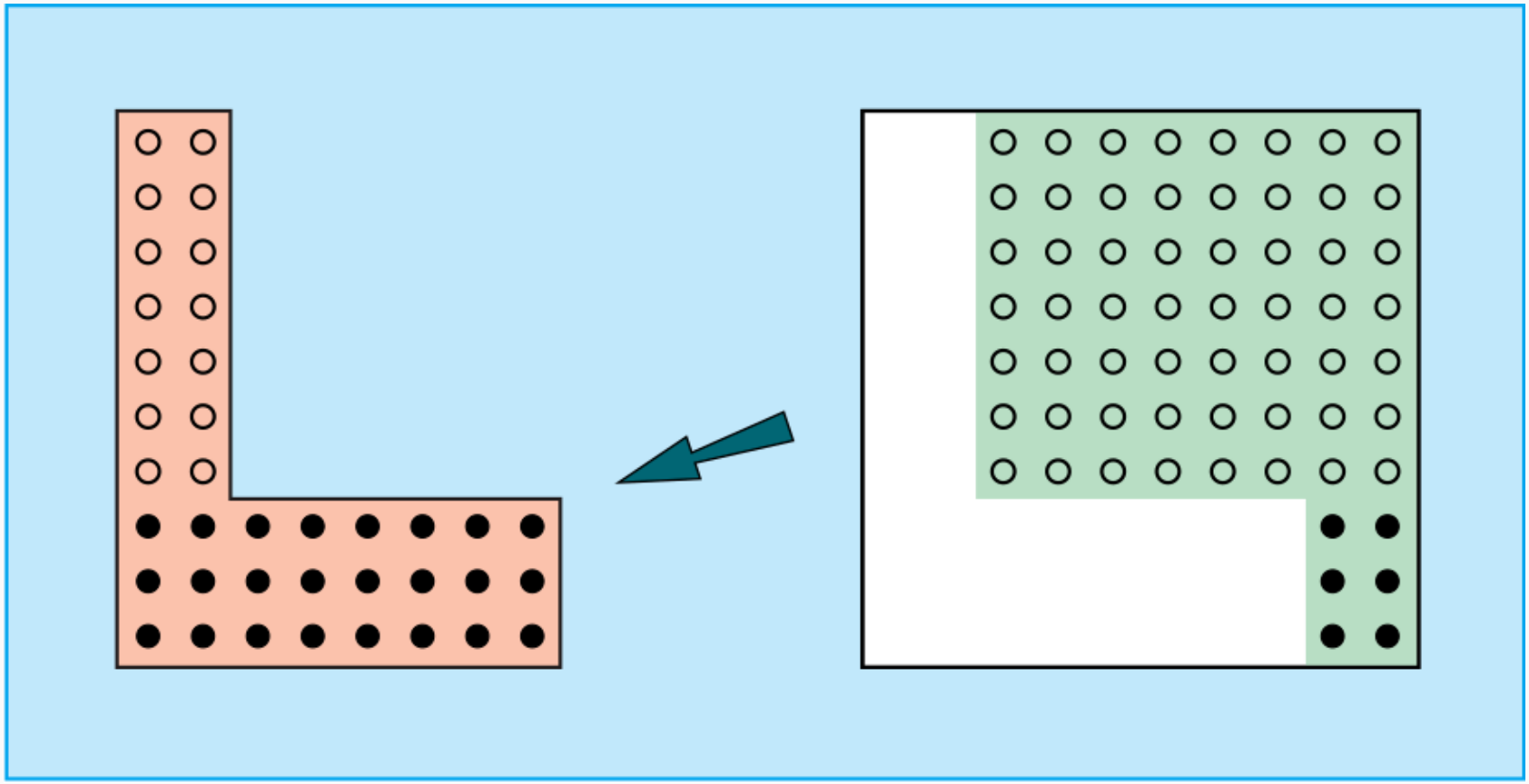
# Sensitivity: 24/30 = 80%



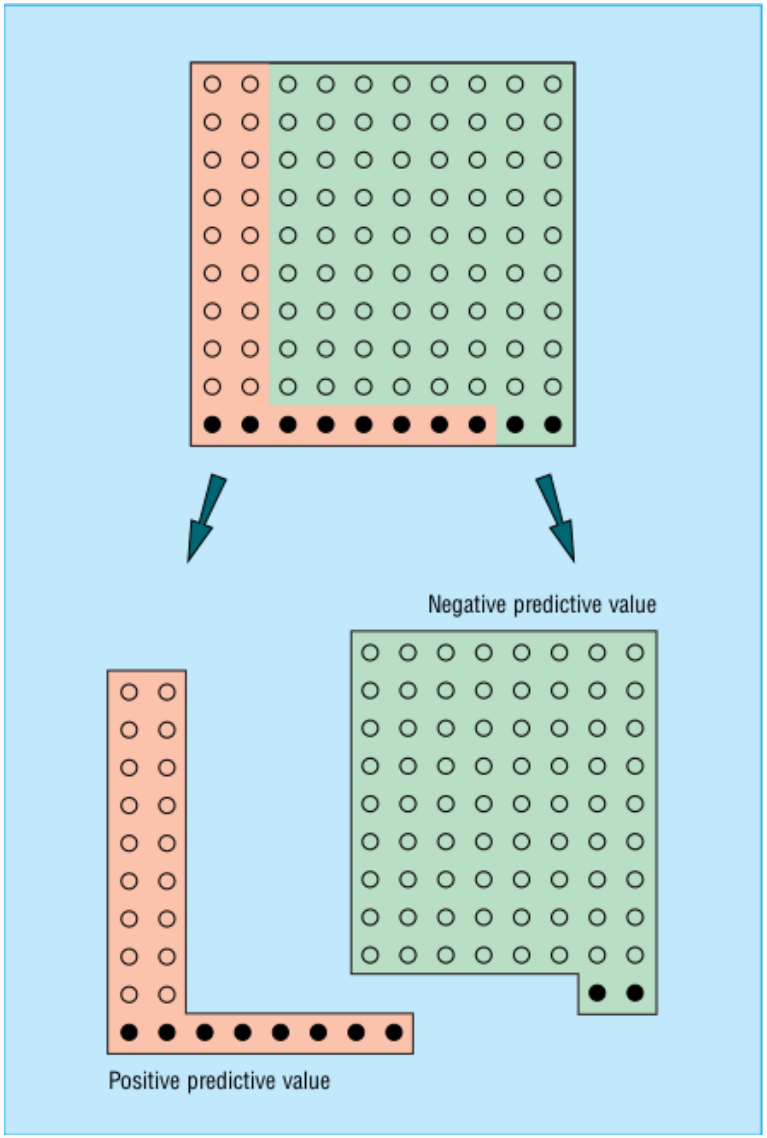
# Specificity: 56/70 = 80%



# Positive Predictive Value: 24/38 = 63%



# Lower Prevalence: PPV = 31%



# An example: Vollenbrock *et al*, BSJ (2019)

- A recent study compared the diagnostic performance of T2-weighted MRI only vs. T2W and diffusion-weighted MRI together for the assessment of residual tumour in patients who underwent chemotherapy for oesophageal cancer.

| | Tumour | No Tumour |
|---|---|---|
| +ive Test | 37 | 6 |
| -ive Test | 2 | 6 |

- Sensitivity: $37/(37 + 2) = 95\%$
- Specificity: $6/(6 + 6) = 50\%$
- Positive Predictive Value: $37/(37 + 6) = 86\%$

# Survival Analysis

# Definition

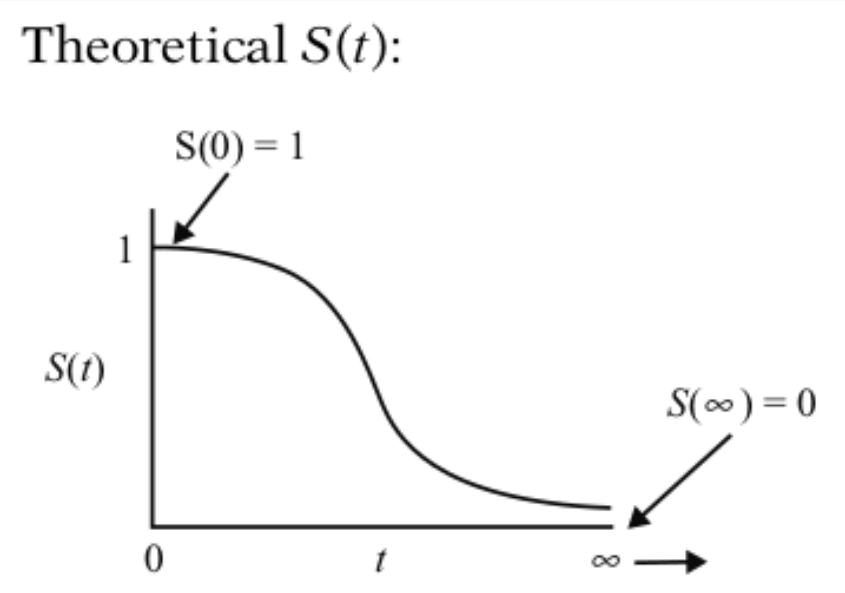
- **Survival analysis** is a set of methods and procedures for the analysis of data where the outcome of interest is the *time until an event of interest occurs*.
- The **event** of interest can be death, disease onset, relapse, etc.
- We typically consider only one event, but we can also study **recurrent events** (e.g. fractures, hospitalizations) or **competing events** (e.g. death from cancer vs. other causes).

# Key Difference: Censoring

- We are not always able to observe the event; in this case, we say the observation is **censored**.
- This can be due to the study ending, loss to follow-up, or competing event.
- We need to take censoring into account in order to perform unbiased inference.

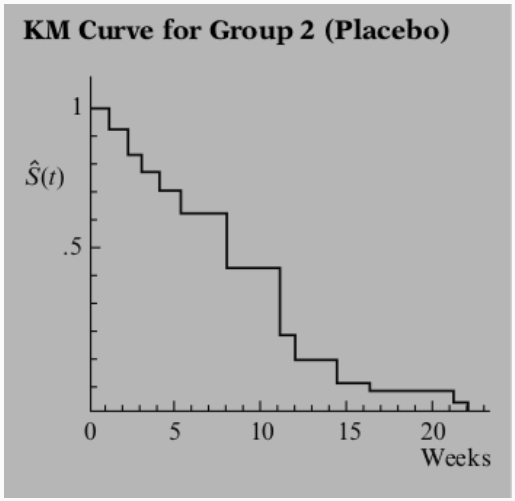
# Survival function

- The main quantity of interest is the **survival function**
- $S(t)$ is the probability of "surviving" past time $t$, i.e. that the event of interest will occur after time $t$.



# Kaplan-Meier Estimator

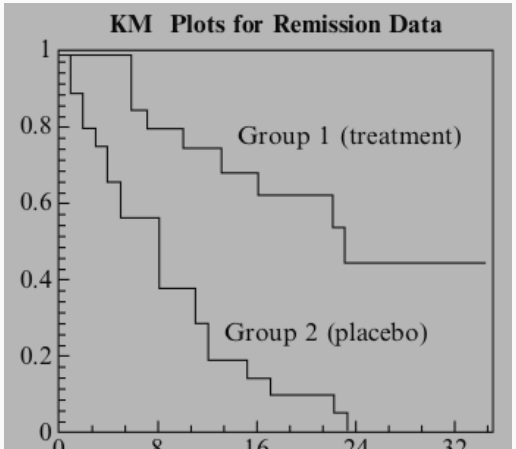
- A very popular way of estimating the survival function from observed data is the **Kaplan-Meier method**.
- It is a nonparametric estimator (i.e. there is no assumption on the distribution of failure times).
  - This leads to a "step function" for the estimate.




- A confidence band around the estimated survival function can be obtained using **Greenwood's formula**.

# Log-Rank Test

- If the observations are grouped into a treatment and a control group, we may be interested in comparing the survival functions of the two groups.
  - If the treatment is effective, the survival function should lie *above* that for the control group.
- The **log-rank test** can be used to obtain a p-value under the null hypothesis that both groups have the same survival function.
  - But this only provides a p-value, no confidence interval for the absolute risk difference.



# Confounding in Survival Analysis

- As above, there are three ways to adjust the inference for confounding:
  1. Randomisation;
  2. Stratification (including Regression);
  3. Weighting.
- Stratified Kaplan-Meier can be used when there are is small number of strata.
- **Cox regression** and **Accelerated Failure Time Models** can be used more generally (but Cox requires an assumption of proportional hazards).

# References

# References i


- Adeyeye, T. E., Yeung, E. H., McLain, A. C., Lin, S., Lawrence, D. A., & Bell, E. M. (2018). Wheeze and Food Allergies in Children Born via Cesarean Delivery: The Upstate KIDS Study. *American Journal of Epidemiology*, 188(2), 355-362.
- Hanley, D. F., Thompson, R. E., Rosenblum, M., Yenokyan, G., Lane, K., McBee, N., ... & Ullman, N. (2019). Efficacy and safety of minimally invasive surgery with thrombolysis in intracerebral haemorrhage evacuation (MISTIE III): a randomised, controlled, open-label, blinded endpoint phase 3 trial. To appear in *The Lancet*.

# References ii


- Kleinbaum, D. G. (2010). *Survival analysis, a self-learning text*. New York: Springer.
- Loong, T. W. (2003). Understanding sensitivity and specificity with the right side of the brain. *British Medical Journal*, 327(7417), 716-719.
- Vollenbrock, S. E., Voncken, F. E. M., van Dieren, J. M., Lambregts, D. M. J., Maas, M., Meijer, G. J., ... & Ter Beek, L. C. (2019). Diagnostic performance of MRI for assessment of response to neoadjuvant chemoradiotherapy in oesophageal cancer. To appear in the *British Journal of Surgery*.

# Appendix

# Regression Cheat Sheet

| Regression Type | Outcome Type | Example |
|---|---|---|
| Logistic Regression | Binary | Disease or not |
| Binomial Regression | Proportion | Readmission per unit |
| Poisson Regression | Count | # SSI per surgeon |
| Linear Regression | Continuous | Blood pressure |
| Cox Regression | Time to Event | Time to death |

# Further questions?