PHPS20010 Biostatistics Tutorial Exercises with Answers

1. DESCRIPTIVE STATISTICS

Read the following table of descriptive statistics carefully, and consider the following points for each of the three measurements:

i. Appropriateness of descriptive statistics used and if not appropriate, what statistics would be better
ii. Suggest a likely distribution (probability curve) of the measurements?
iii. Do the statistics presented let you deduce anything about the numbers of people with particular values or categories in the sample, or in the population?

Patient characteristics	Placebo	Treatment
Age, years	Mean: 76 SD: 12	Mean: 77 SD: 11
Sex (female)	Median: 1 IQR: 0, 1	Median: 1 IQR: 0, 1
Length of ICU stay, days	Mean: 5 SD: 3	Mean: 2 SD: 4

SD: Standard Deviation; ICU: Intensive Care Unit

Answers:

Age

- Numerical / continuous and likely Normally distributed
- Mean and SD are OK.
- Can deduce it's an older population. If Age is Normal 95% of the sample are between their mid-50s and 100 years of age (approx. ±2 SD).

Female Gender

- Inappropriate summary statistics: median and IQR are used for numerical measurements
- Categorical variable should be described using N and %.
- If we had that information, we could work out the percentage (and the number) of males. As it is, we can say nothing.

Length of ICU stay

- Time durations and counts tend to be right-skewed (see: 2 SDs below mean is impossible)
- A mean therefore is probably an overestimate.
- Should use median, and IQR.

PHPS20010 Biostatistics Tutorial Exercises with Answers

2. RISK, UNCERTAINTY, AND COMMUNICATION

Consider the following scenarios for hypothetical new infectious diseases (in a fixed time period):

Disease	A	B	C	D	E	F
Prevalence	0.01	0.01	0.05	0.05	0.2	0.2
Mortality risk	0.12%	40%	0.12%	8%	0.12%	5%
Risk Difference*	0.04%	39.9%	0.04%	7.9%	0.04%	4.9%
Relative risk of mortality	1.5	500	1.5	100	1.5	62.5
* Versus baseline population mortality of 0.0008						

The Population Attributable Risk of mortality represents the contribution of a disease (or any risk factor) to the total population mortality per time period. It can be calculated from the risk difference (aka attributable risk or excess risk) between those with and without disease, multiplied by the prevalence. It is usually expressed out of 100,000 population by multiplying by 100,000.

i. Which of these is the most critical disease on a population level, and which is most critical for a patient presenting to you with a diagnosis?
ii. For the disease likely to have most impact on a population level, speculate about how you advise a patient who has tested positive?

Answers:

i. The PARs are:

PAR (per 100,000)	0.4	399.2	2	396	8	984

- The worst burden on the population is F
- About 1% of the population will die from such a prevalent disease.
- The worst for an individual is B with a 40% mortality for this patient.

ii. Advising patient with disease F:

- Emphasize high relative risk: 62.5-fold
- The patient needs to understand the risk to the rest of the population as well as their own health – PAR implies 1% of population will die (~50,000 people in Ireland)
- High prevalence of F implies high infectivity/virulence, consider measures to reduce spread?

PHPS20010 Biostatistics Tutorial Exercises with Answers

3. CONFIDENCE INTERVALS

A Mean, and 95% Confidence intervals for blood pressure measurements were calculated for a sample of 50-year old women as given below. The statistician provided the confidence limits, but neglected to label them.

Mean Systolic blood pressure pre-treatment = 118.0	(96.1, 107.9)
Mean Systolic blood pressure post-treatment = 121.0	(115.9, 126.1)
Mean Arterial Pressure = 102.0	(114.2, 121.8)

i. Assign the confidence interval to the correct variable in the table
ii. Was the systolic blood pressure more variable pre-treatment or post-treatment?
iii. The researcher requested 99% confidence intervals, rather than 95%. What effect will that have on the width of the interval?

Answers:

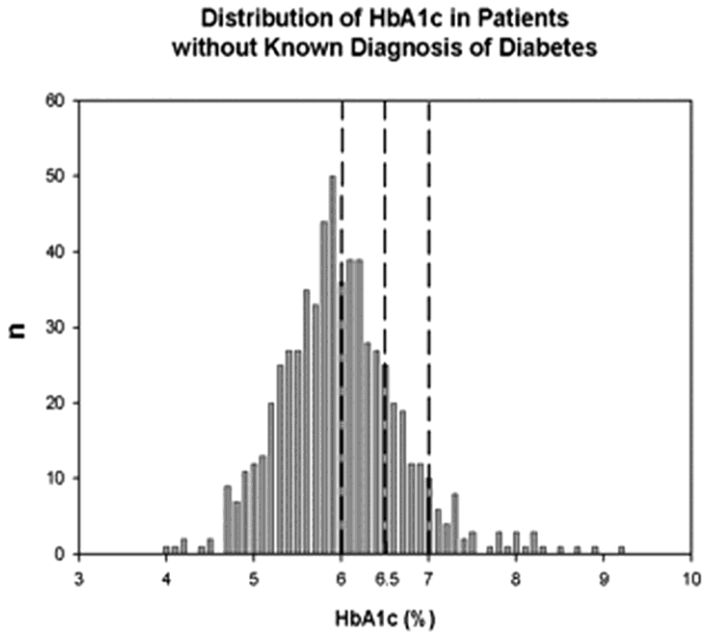
i. Calculate the midpoint of the confidence intervals:
SBP pre-treatment with a mean of 118 has a 95% CI centred on 118: **114.2, 121.8**
SBP post Treatment has a higher mean, and CI centred on that **115.9, 126.1**
Mean Arterial Pressure has a lower mean, and the CI is shifted lower **96.1, 107.9**

ii. The Post-treatment SBP is more variable because the Confidence Interval width is 126.1-115.9 = 10.2, versus 7.6
This assumes the sample size is the same pre and post.

iii. The Confidence Interval widens at a higher level of confidence.
A higher confidence that you're covering the true population value is only obtained by making the interval wider

4. DISTRIBUTIONS AND HYPOTHESIS TESTING

Glycated haemoglobin A1c (HbA1C) is used as a test for pre-diabetes or diabetes, as it reflects long-term blood sugar levels. In a study* of 971 hospital in-patient admissions (unselected) the following distribution was observed:



The mean HbA1C was 6.02, with a standard deviation of 0.87. Thresholds for further testing (6.0), for provisional diabetes diagnosis (6.5) and for treatment of diabetes (7.0) are shown on the graph.

* https://doi.org/10.1210/jc.2009-1151

i. Assuming a Normal distribution, ***roughly*** what percentage of this patient population would require at least further testing?

- Normal distribution is symmetrical
- The mean is near 6.0, so approx. 50% of the population should warrant further action
- From the paper: "*…323 of 629 NKD patients (51%) had at least one HbA1c of at least 6.0%...".*

ii. Comparison of HbA1C between patients with and without known diabetes gave a p-value <0.001. What statistical test should have been employed to obtain that p-value? Interpret the p-value.

- To compare the means, usually done by a ***t test***.
- The p-value is very small, implying the data is not compatible with the means being the same
- At a traditional 5% level it is a "statistically significant" difference between the two groups.

iii. The paper states:

"… The remaining 1132 admissions, representing 971 unique patients, were then categorized into "known diabetics" and "NKD" [no known diabetes] *as defined by the above criteria."*

"… To account for the correlated nature of these data, generalized estimating equations were used to assess associations between demographic and clinical characteristics by diabetes status and by HbA1c."

Explain in simple terms why they did not use the most obvious statistical test.

- More HbA1C measurements than patients implies ***repeated measures on some patients***.
- Non-independent data points
- Usual t test is not valid
- The specified method – Generalized Estimating Equations – is a regression model that accounts for this non-independence / correlation within a patient's repeated measures (not expected to recall this method)