STAT115: Introduction to Biostatistics

University of Otago
Ōtākou Whakaihu Waka

Lecture 12: Sampling Distributions

Outline

- Previous:
  - Introduction to statistical modelling
  - Looked into the normal distribution
- Today:
  - Look at sampling distribution
  - Explore: how precise is the estimate $\bar{y}$?

Example

- Previously we have been exploring cholesterol in heart attack patients
- Today we will use a different example
- Data from urine tests of $n = 314$ children (aged 0 – 17 years)
  - (log) GAG concentration[1]
  - GAG: glycosaminoglycan
    - Test is used to diagnose disorders of glycosaminoglycan metabolism
    - Glycosaminoglycans are important in cell signalling
- Data were collected to help paediatricians assess normal level of GAG concentration
- Today we'll consider a simpler problem
  - What is the expected (or mean) GAG concentration?

[1] We will refer to this as the concentration from here on

Data

- The data are in `GAG.csv`
- Import the data into R [2]

```
GAG = read.csv('GAG.csv')
```

- The function `head` shows us the first few lines of data

```
head(GAG)
```

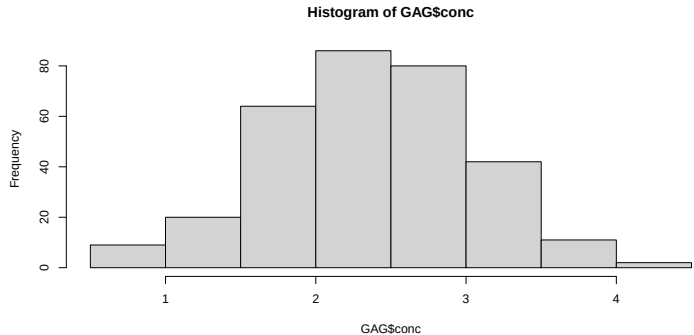
```
##    age  conc
## 1 0.00 3.135
## 2 0.00 3.170
## 3 0.00 2.827
## 4 0.00 2.923
## 5 0.01 2.885
## 6 0.01 3.254
```

[2] Recall there are several ways to do this: see week 1 of lectures

Data

- Look at a histogram

```
hist(GAG$conc) # dollar sign: selects the appropriate variable (conc)
```



- We can 'adapt' this plot to change axes labels, title, etc.
  - Keep it simple, getting an idea of the data

Recap: normal model

- We model the data as from a normal distribution
  - Modelling GAG concentration as being normally distributed
- Two parameters μ and σ
- Parameters are unknown
  - μ: mean GAG concentration
  - σ: standard deviation of GAG concentrations
- Return to our question: what is the expected (or mean) GAG concentration?
  - Estimate μ with sample mean
  - $\hat{\mu} = \bar{y}$

```
ybar_conc = mean(GAG$conc)
ybar_conc
```

```
## [1] 2.364
```

Critical thinking

- Do we now know the expected GAG concentration?
  - That we could use (if we were a paediatrician) seeing patients

Critical thinking

- Do we now know the expected GAG concentration?
  - That we could use (if we were a paediatrician) seeing patients
- No, we don't
  - Mean GAG concentration is a parameter μ
  - Estimated it with a statistic: sample mean, $\bar{y}$
- How precise is the estimate?
  - If we took another sample of 314 children, how much would the estimate change?
  - Would you 'trust' the estimate more, less, or the same, if:
    - The estimate was from a sample of 8 children?
    - The estimate was from a sample of 50 000 children?

Thought experiment

- How close to μ is $\bar{y}$?

Thought experiment

- How close to μ is $\bar{y}$?
- To answer it, let's play god:
  - Assume that GAG concentration really is normal
  - Pretend that we know μ and σ
    - $\mu = 2.4$
    - $\sigma = 0.75$
- Take a sample of size $n = 314$ from the population
  - Observe how close the sample mean $\bar{y}$ is to μ
- Take many (separate) samples of size n
  - See how much $\bar{y}$ varies from one sample to another

Let's try it

- We saw a function previously for simulating from a normal distribution

```
rnorm(n,mean,sd)
```

- Generates a sample of size `n` from a normal distribution with mean (`mean`) and std deviation (`sd`)

```
n = 314; mu = 2.45; sigma = 0.75
y = rnorm(n = n, mean = mu, sd = sigma)
mean(y)
## [1] 2.52
```

- True mean: $\mu = 2.45$; sample mean: $\bar{y} = 2.52$

What if we took a lot of samples?

- Repeat this m times (using R)
  - You will not be expected to replicate the R code below

```
m = 10000 # the number of samples
ybar = rep(NA, m) # this 'initializes' a vector to store each
# of the m sample means
for(i in 1:m){ # repeats the code below m times
  y = rnorm(n, mu, sigma) # takes a sample of size n = 314
  ybar[i] = mean(y) # finds the sample mean and stores it in ybar
}
```

What if we took a lot of samples?

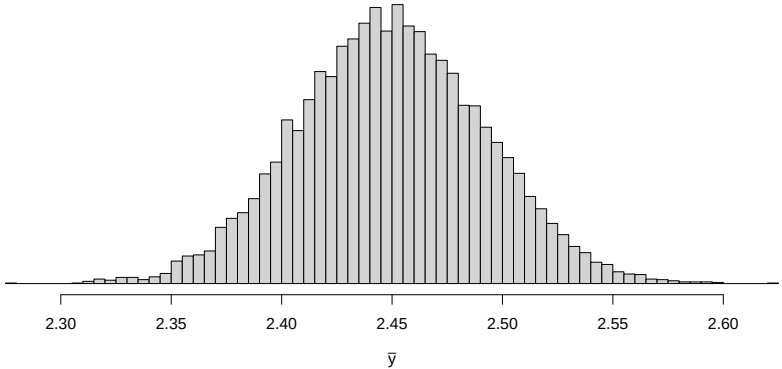
- The first few sample means are:

```
head(ybar)
```

```
## [1] 2.48 2.46 2.45 2.46 2.44 2.50
```

- We could look at a histogram of these
  - Get an idea of the distribution of sample means
  - Evaluate how variable $\bar{y}$ is: one sample to another
  - Assess whether $\bar{y}$ accurately estimates the mean (on average)

What if we took a lot of samples?



Sampling distribution

- This is called the sampling distribution
  - Sampling distribution of $\bar{y}$
- Tells us how we would expect our statistic ($\bar{y}$) to vary from one sample to another
- From the histogram we can see
  - On average it is 2.45: the value of μ
  - Sample means less than 2.35 or larger than 2.55 are unlikely

What if?

- We can use this to answer 'what if' questions, e.g.
- What is the chance of observing a sample mean as extreme as $\bar{y} = 2.36$
  - If the $\mu = 2.45$ and $\sigma = 0.75$?
- Look at the histogram again:
  - Possible, but unlikely
- Could use R to count how many samples (of 10 000) had mean less than 2.36
  - Estimate the probability
  - R shown for interest only

```
sum(ybar < ybar_conc) # ybar_conc = 2.36 (from data)
```

```
## [1] 218
```

What is extreme?

- We asked ‘what is the chance of observing a sample mean as extreme as ...’
  - Did we answer that correctly?

What is extreme?

- We asked 'what is the chance of observing a sample mean as extreme as ...'
  - Did we answer that correctly?
- No: we looked at chance of observing a sample mean less than 2.36
  - A sample mean higher than 2.54 is just as extreme as one below 2.36
  - Both are 0.09 units away from the true mean ($\mu = 2.45$)
- An extreme observation could be below or above the mean
  - Calculating the probability of an extreme value needs to account for both
- This is a principle we will use often

Theory

- It turns out that when we have a normal model for y
  - The sampling distribution (distribution of sample means $\bar{y}$) is also normally distributed
- What are the mean and variance?
  - The mean of the sampling distribution is μ
  - The variance of the sampling distribution is $\frac{\sigma^2}{n}$
  - The standard deviation of the sampling distribution is $\frac{\sigma}{\sqrt{n}}$

Theory

- Where do these results come from?
  - We worked these out a few lectures ago! (copied below)
    - The expected value of the sample mean is

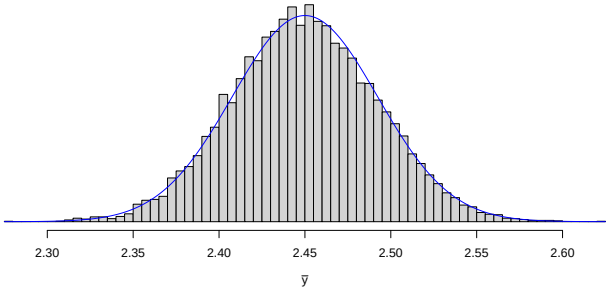
$$E\left[\frac{Y_1 + Y_2 + \ldots + Y_n}{n}\right] = \frac{1}{n}E[Y_1] + \frac{1}{n}E[Y_2] + \ldots + \frac{1}{n}E[Y_n]$$
$$= \mu$$

    - The variance of the sample mean is

$$Var\left(\frac{Y_1 + Y_2 + \ldots + Y_n}{n}\right) = \frac{1}{n^2}Var(Y_1) + \frac{1}{n^2}Var(Y_2) + \ldots + \frac{1}{n^2}Var(Y_n)$$
$$= \frac{\sigma^2}{n}$$

Sampling distribution

- When using a normal model for y, the sampling distribution for $\bar{y}$
  - Normally distributed with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$
- For the example above, the sampling distribution has:
  - Mean: 2.45, standard deviation $\frac{0.75}{\sqrt{314}}$
- Compare to the sampling distribution found in R



Sampling distribution

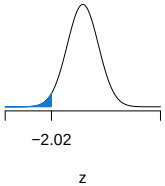
- Use our knowledge of the normal distribution to earlier questions
- What is the chance of observing a sample mean as extreme as $\bar{y} = 2.36$?
  - If the $\mu = 2.45$ and $\sigma = 0.75$?
- Three steps
  1. Find mean and sd of sampling distribution
  2. Convert to z-value
  3. Find the probability

Sampling distribution

- Mean: $\mu = 2.45$
- Standard deviation: $\frac{\sigma}{\sqrt{n}} = \frac{0.75}{\sqrt{314}}$
- z-value: $z = \frac{\bar{y}-\mu}{\frac{\sigma}{\sqrt{n}}} = \frac{2.36-2.45}{\frac{0.75}{\sqrt{314}}} = -2.021$
- Probability of z-value less than -2.021

```
z = (ybar_conc - mu) / (sigma / sqrt(n)) # z-value
pnorm(z)
```

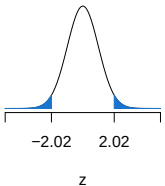
```
## [1] 0.0216
```



- Probability of z-value more extreme than -2.021

```
2*pnorm(z) # same area in each tail (see graphic)
```

```
## [1] 0.0432
```



Does this make sense?

- The standard deviation of the sampling distribution $\frac{\sigma}{\sqrt{n}}$
  - Decreases as n increases
- Makes sense
  - As the sample size (n) increases, the estimate $\bar{y}$ is increasingly precise
- If n is small ($n = 1$)
  - Sample mean is the same as an observation: same sd (σ)
- If n is large ($n = 1\,000\,000$)
  - Standard deviation of the sample mean is $1/1000$th the sd of observations
  - Lots of data: sample mean is a precise estimate of true mean

Summary

- Introduced the concept of sampling distribution
  - Tells us how much $\bar{y}$ varies from one sample to the next
- Introduced some core principles that we will see again and again
- Standard deviation of sampling distribution is $\frac{\sigma}{\sqrt{n}}$
  - Use this to evaluate how precise an estimate is
  - Problem: relies on σ being known
  - What happens if σ is unknown
    - Always the case in the real world
  - Explore in the next lecture