STAT 110: Week 4

University of Otago

Outline

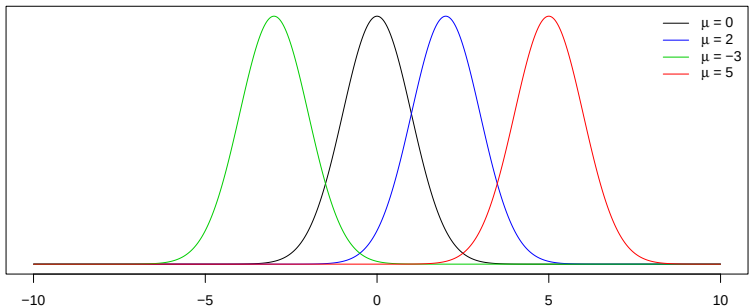
- Previous lectures:
  - Introduction to probability, random variables
  - First example of a statistical model
    - Normal model
- Today: learn more about the normal distribution

Normal distribution

- We used a normal model to describe flipper length (gentoo penguins in Palmer archipelago)
- Is the normal model appropriate
  - Does it make sense scientifically
    - Understand 'properties' of a normal distribution
    - Looked at some aspects in last lecture
    - Understand more about the normal distribution today
  - After estimation: check model fit
    - Looked briefly at this in last lecture
    - Consider it further in future lectures

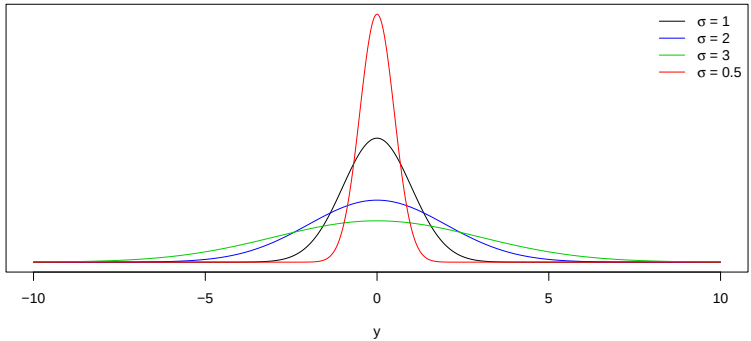
Recap: normal distribution

- Described by two parameters
  - Mean μ
  - Standard deviation σ
- Changing μ shifts the pdf side to side



Recap: normal distribution

- Described by two parameters
  - Mean μ
  - Standard deviation σ
- Changing σ compresses or expands the pdf



IQ scores

- IQ tests are designed so that scores are (approximately) normally distributed
  - $\mu = 100$
  - $\sigma = 15$
- We may be interested in knowing things like:
  - What is the probability of a randomly chosen individual scoring less than 85?
  - What is the probability of a randomly chosen individual scoring between 85 and 115?
  - For membership Mensa require a score at or above the 98th percentile on certain standardized IQ tests. For an IQ test (as above) what score would you need?
- All of these require us to be able to find probabilities from the normal distribution

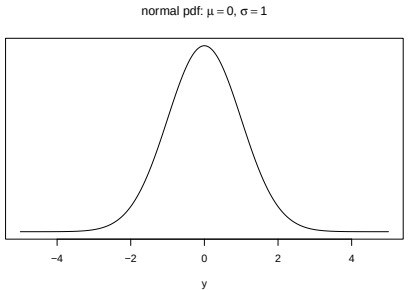
Probabilities

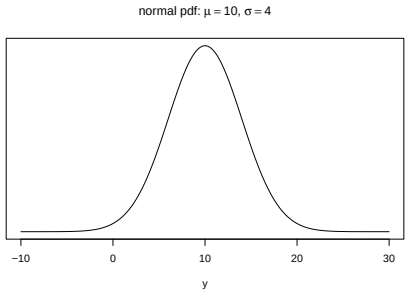
- Recall: we find probabilities by finding the area under pdf
- The normal pdf is a mathematical function: $f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y-\mu)^2\right)$
  - Not expected (or required) to remember or understand this
  - Mathematical representation of the pdfs we saw in earlier slides
- Theory: to find probabilities we can use calculus and integrate $f(y)$ [1]
  - Problem: can't integrate $f(y)$ by hand
- Historical solution: tables of values we could refer to
  - Problem: lots of possible values of μ and σ
  - Solution: find them for a single standardized version of the distribution

[1] Integration can be thought of as (mathematically) finding the area under curve

Standard normal distribution

- Normal pdfs have the same shape
  - Irrespective of the value of μ, σ
    - Hard to see on the previous plots
    - More clear if change the scale of the axes for different values of μ, σ





- Idea: work with a standard normal distribution: $\mu = 0$, $\sigma = 1$

Standardizing

- Idea: define a standard normal distribution
  - $\mu = 0$, $\sigma = 1$
- Find probabilities, etc, for this standard distribution
- Convert a value (y) to a z-score
  - y-value from distribution with mean μ and standard deviation σ
  - z-score from distribution with mean 0 and standard deviation 1
  - Going from y to z is often called standardizing
- The z-score tells us how many standard deviations above the mean a value is
  - $z = 1$: value is 1 standard deviation above the mean
  - $z = -1.5$: value is 1.5 standard deviations below the mean

Standardizing

- We can find a z-score from y

$$z = \frac{\text{value} - \text{mean}}{\text{standard deviation}} = \frac{y - \mu}{\sigma}$$

- IQ test of $y = 115$:

$$z = \frac{y - \mu}{\sigma} = \frac{115 - 100}{15} = 1$$

  - An IQ test of 115 is one standard deviation above the mean

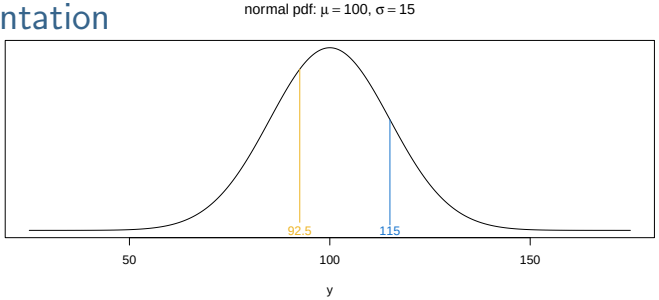
- We can also find y from a z-score

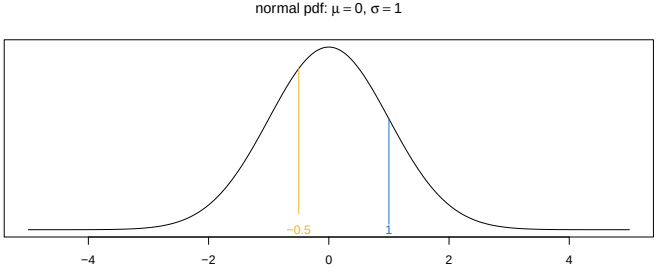
$$y = \mu + z\sigma$$

- A z-score of 1 for IQ corresponds to a score of:

$$y = 100 + 1 \times 15 = 115$$

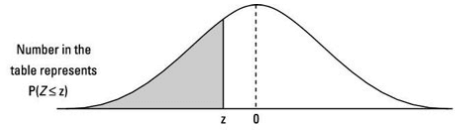
Graphical representation





Finding probabilities: deep dark past

- We used to find probabilities from tables



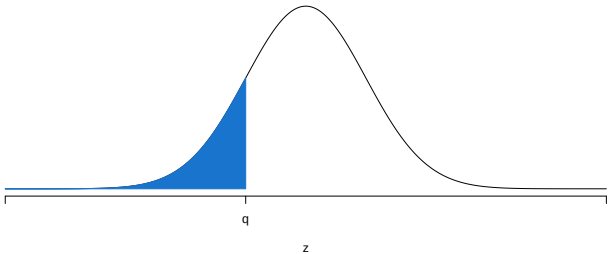
z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
–3.6	.0002	.0002	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001
–3.5	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002
–3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
–3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
–3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
–3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
–3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
–2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
–2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
–2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
–2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
–2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
–2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
–2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
–2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
–2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
–2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183

Finding probabilities: computing age

- We can find them using a graphical calculator or computer
- We will use R
- R has four functions for the normal distribution
  - `dnorm`: **d**ensity function
  - `pnorm`: **p**robability function
  - `qnorm`: **q**uantile function
  - `rnorm`: generate **r**andom values
- In STAT 110, most our interest is in `pnorm` and `qnorm`
  - Look at each in turn

Probability function

- This is best seen graphically
- The blue area is given by `pnorm(q)`
  - $\Pr(Z < q)$



- Look at three examples

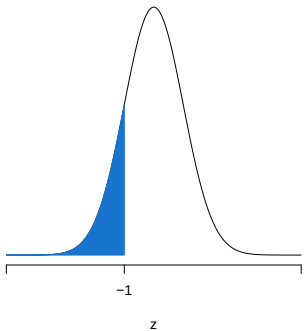
Example 1

- What is the probability that IQ is less than 85?
- Find z-score:

$$z = \frac{y - \mu}{\sigma} = \frac{85 - 100}{15} = -1$$

- Find $\Pr(Z < -1)$

```
mu = 100; sigma = 15 # the mean and sd for IQ
z = (85 - mu)/sigma # finding the z-score
pnorm(z)
## [1] 0.159
pnorm(-1) # for those who want to check
## [1] 0.159
```



Example 2

- Probability that IQ is more than 120?

$$z = \frac{y - \mu}{\sigma} = \frac{120 - 100}{15} = 1.3333$$

- Use `pnorm` to find $\Pr(Z < 1.3333)$ (gold area)

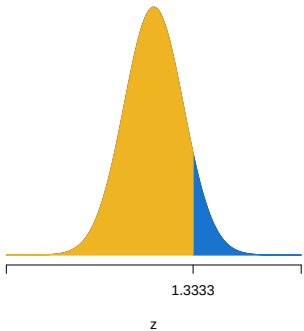
```
z = (120 - mu)/sigma # finding the z-score
pnorm(z)
```

```
## [1] 0.909
```

- $\Pr(Z > 1.3333)$ (blue area) is the complement
  - $\Pr(Z > z) = 1 - \Pr(Z < z)$

```
1-pnorm(z)
```

```
## [1] 0.0912
```



Example 3

- Probability that IQ is between 110 and 130?

$$z_{110} = \frac{y - \mu}{\sigma} = \frac{110 - 100}{15} = 0.6667$$

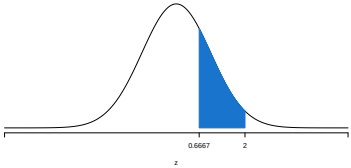
$$z_{130} = \frac{y - \mu}{\sigma} = \frac{130 - 100}{15} = 2$$

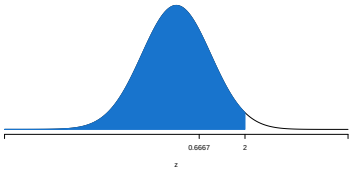
```
z110 = (110 - mu)/sigma # finding the z-score
z130 = (130 - mu)/sigma
```

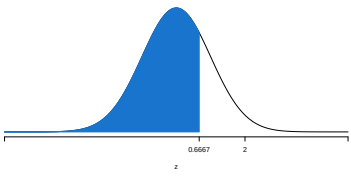
- $\Pr(z_{110} < Z < z_{130}) = \Pr(Z < z_{130}) - \Pr(Z < z_{110})$
  - Best seen graphically on RHS

```
pnorm(z130)-pnorm(z110)
```

```
## [1] 0.23
```





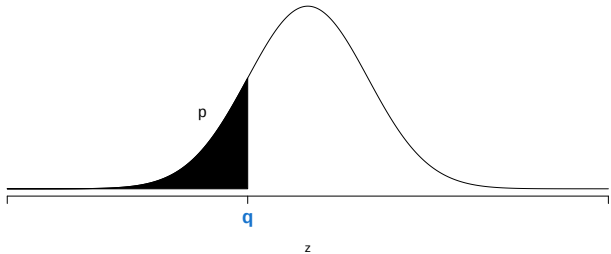


Important properties

- We can use this to learn some important characteristics of a normal distribution
- $\Pr(-1 < Z < 1) = 0.683$
  - Approximately 68% of values should be within 1 sd of the mean
- $\Pr(-2 < Z < 2) = 0.955$
  - Approximately 95% of values should be within 2 sd of the mean
- $\Pr(-3 < Z < 3) = 0.997$
  - More than 99% of values should be within 3 sd of the mean
- Challenge: confirm these numbers using `pnorm` in R before next class

Quantile function

- Basically the same graphic as before: interest is switched
- The value q is given by `qnorm(p)`
  - The value of p is the black area (known)



- Look at an example

Example

- What score is required for Mensa membership
  - At or above the 98th percentile
    - In the top 2%
- Find the z-score corresponding to $p = 0.98$

```
z = qnorm(0.98)
```

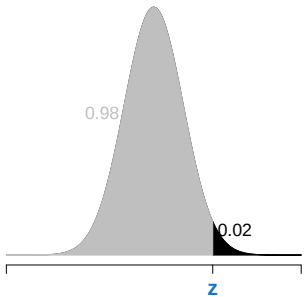
- Find the y-value

$$y = \mu + z\sigma$$

```
mu + z * sigma

## [1] 131
```

- Need an IQ score of 131 or higher



z or y?

- Throughout we have done calculations using standard normal
  - Standardized to find z
- With R it is comparatively easy to find using y
  - `pnorm` has optional arguments for the mean and sd
- First example: $\Pr(IQ < 85)$

```
pnorm(q = 85, mean = 100, sd = 15)
```

```
## [1] 0.159
```

- Rstudio guides you as to the arguments (in R)
- Important to know about z / standardization
  - Required knowledge in the scientific world
  - Need it to understand how confidence interval and t-tests work

Summary

- Looked in some detail at normal distribution
  - Standardization and z-scores
  - Finding probabilities from z-scores
  - Finding z-scores from probabilities
- Next class: sampling distributions
  - If we took another sample, how much variation would we expect in the sample mean $\bar{y}$?

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 flipper length of gentoo penguins
- Today we will use a different example
- Data from urine tests of $n = 314$ children (aged 0 – 17 years)
  - (log) GAG concentration[2]
  - 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?

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

Data

- The data are in `lect4GAG.csv`
- Import the data into R [3]

```
lect4GAG = read.csv('lect4GAG.csv')
```

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

```
head(lect4GAG)
```

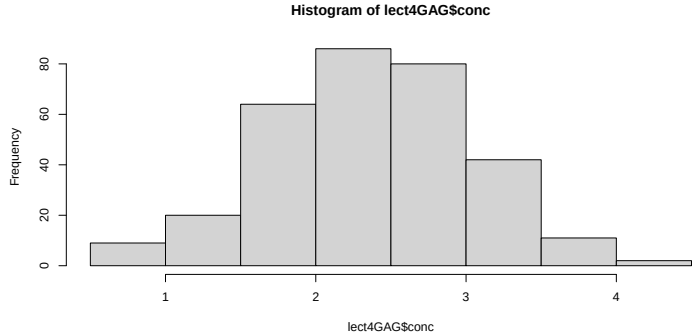
```
##    age conc
## 1 0.00 3.14
## 2 0.00 3.17
## 3 0.00 2.83
## 4 0.00 2.92
## 5 0.01 2.88
## 6 0.01 3.25
```

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

Data

- Look at a histogram

```
hist(lect4GAG$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(lect4GAG$conc)
ybar_conc
```

```
## [1] 2.36
```

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?

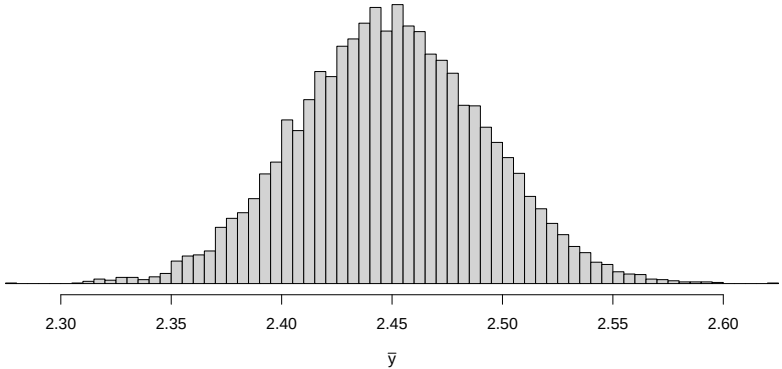
- The first few sample means are:

```
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! (lecture 8; copied below)
    - The expected value of the sample mean is

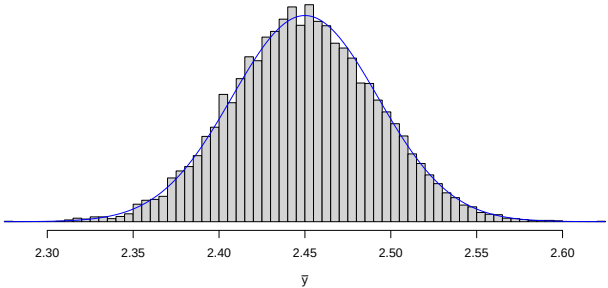
$$E\left[\frac{Y_1 + Y_2 + \ldots + Y_n}{n}\right] = \frac{1}{n}E[Y_1] + \frac{1}{n}E[Y_2] + \ldots + \frac{1}{n}E[Y_n]$$
$$= \mu$$

    - 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

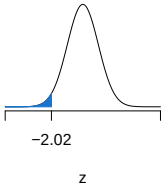
- Use our knowledge of the normal distribution to earlier questions
- What is the chance of observing a sample mean as extreme as $\bar{y} = 2.36$?
  - If the $\mu = 2.45$ and $\sigma = 0.75$?
- Three steps
  1. Find mean and sd of sampling distribution
  2. Convert to z-value
  3. Find the probability

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)
```

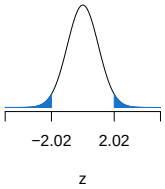
```
## [1] 0.0216
```



- 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

Outline

- Previous:
  - Introduction to (normal) statistical model
  - Sampling distributions
    - Describe variation in the sample mean $\bar{y}$ (or any other statistic) from one sample to another
    - Relies on us knowing σ
- Today:
  - Use that to find confidence interval
    - Interval estimate for the parameter value
  - Look at what happens when σ is unknown

Example

- Continue using the GAG concentration data
  - Data from urine tests of $n = 314$ children (aged 0 – 17 years)
  - (log) concentration of glycosaminoglycan (GAG)
- Asking: what is the expected (or mean) GAG concentration?

Sampling distribution

- Recall we have a normal model for the data
  - Data come from a normal distribution with mean μ and standard deviation σ
- Last lecture we found the sampling distribution for $\bar{y}$
  - Distribution that describes how $\bar{y}$ will vary from one sample to another
  - Sampling distribution is normally distributed (for a normal model)
    - Mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$

Cool result!

- We know about what will happen in repeated samples
  - Without having to take repeated samples!
- If we know the data distribution (i.e. we know μ and σ):
  - We know how variable we expect $\bar{y}$ to be without even sampling from the population
- If we know σ (but don't know μ):
  - Can we use a single sample to tell us about a range of plausible values of μ?

Cool result!

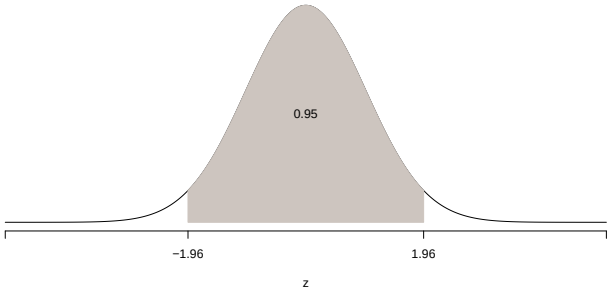
- We know about what will happen in repeated samples
  - Without having to take repeated samples!
- If we know the data distribution (i.e. we know μ and σ):
  - We know how variable we expect $\bar{y}$ to be without even sampling from the population
- If we know σ (but don't know μ):
  - Can we use a single sample to tell us about a range of plausible values of μ?
- Yes!

Excursion: standard error

- Confusing notation to discuss
- Over the past few lectures, we have seen:
  - Population standard deviation σ
  - Sample standard deviation s
  - Standard deviation of sampling distribution of $\bar{y}$
    - It is $\frac{\sigma}{\sqrt{n}}$
    - Has a special name: standard error
    - Can be represented with notation $\sigma_{\bar{y}}$
  - Estimate of the standard deviation of the sampling distribution of $\bar{y}$
    - It is $\frac{s}{\sqrt{n}}$
    - It is often also called the standard error
    - Can be represented with notation $s_{\bar{y}}$
- Very confusing

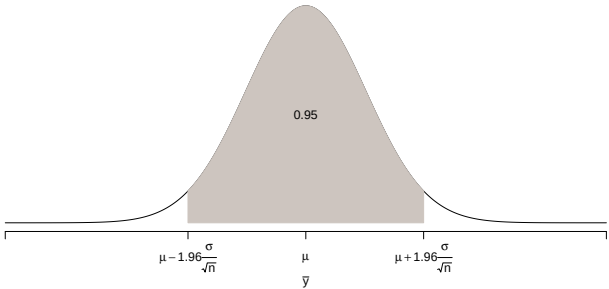
Previous knowledge

- Want to determine an interval estimate of μ from $\bar{y}$
- From our knowledge of normal distribution:
  - 95% of observations will fall within (approx) ± 2 standard deviations of mean
    - More precisely it is ± 1.96
    - In R: `qnorm(0.025)` and `qnorm(0.975)`
  - $\Pr(-1.96 < Z < 1.96) = 0.95$



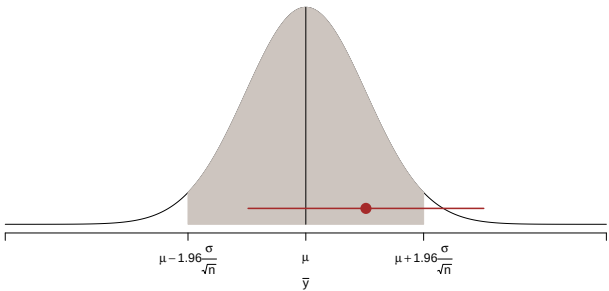
Sampling distribution

- Applying this to the sampling distribution we have:
  - 95% of sample means ($\bar{y}$) are between ± 1.96 standard errors ($\frac{\sigma}{\sqrt{n}}$) of the mean
- 95% of samples we collect will have sample means in the grey area
  - Given by $\mu \pm 1.96\frac{\sigma}{\sqrt{n}}$



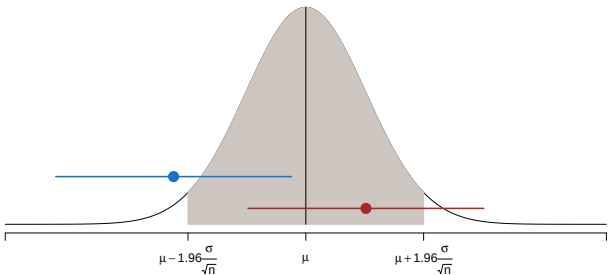
Flipping things I

- Consider any sample mean that is **inside** the shaded grey area
  - We've plotted one in brown on plot below
- Here's the magic:
  - If $\bar{y}$ is inside the grey area ($\mu \pm 1.96\frac{\sigma}{\sqrt{n}}$) (brown point)
  - Then μ (vertical black line) is inside the interval $\bar{y} \pm 1.96\frac{\sigma}{\sqrt{n}}$ (brown interval)



Flipping things II

- Consider any sample mean that is **outside** the shaded grey area
  - We've plotted one in blue on plot below
- Here's the magic:
  - If $\bar{y}$ is outside the grey area ($\mu \pm 1.96\frac{\sigma}{\sqrt{n}}$) (blue point)
  - Then μ (vertical black line) is outside the interval $\bar{y} \pm 1.96\frac{\sigma}{\sqrt{n}}$ (blue interval)



Confidence interval

$$\bar{y} \pm 1.96 \frac{\sigma}{\sqrt{n}}$$

- This is a 95% confidence interval for μ
  - Interval estimate of μ
  - Quantifies how precise the estimate of μ is
- On average, 95% of sample means will lie in shaded grey area (established above)
  - That means that our confidence interval should contain the true μ in 95% of samples
  - Gives us confidence in the procedure (hence the name)
    - Care is needed: we cannot say that there is a probability of 0.95 that μ is in the interval

A few notes on confidence intervals

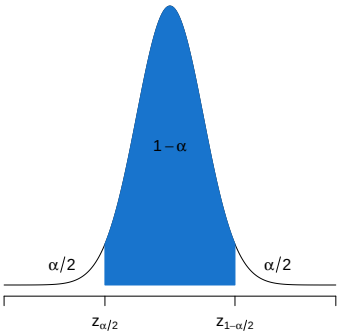
- The confidence interval is in a general form:

$$\text{estimate} \pm \text{multiplier} \times \text{standard error}$$

- estimate: $\bar{y}$
- multiplier:
  - 1.96 for 95% confidence interval
  - More generally, we write $z_{1-\alpha/2}$
    - More details on next slide
- Standard error: $\frac{\sigma}{\sqrt{n}}$

Multiplier

- Multiplier: $z_{1-\alpha/2}$
  - Also referred to as the critical value
- α: significance level
  - significance level = 1 - confidence level
    - 95% interval: $\alpha = 1 - 0.95 = 0.05$
    - 90% interval: what is α?
- $\Pr(Z < z_{1-\alpha/2}) = 1 - \alpha/2$
  - Find z-value so that tails have probability $\alpha/2$



Multiplier

- For a 95% interval
  - $\alpha = 0.05$
  - $1 - \alpha/2 = 0.975$
  - We want to find $z_{0.975}$

```
qnorm(0.975)
```

```
## [1] 1.96
```

- How do we find the multiplier for a 90% interval?

Multiplier

- For a 95% interval
  - $\alpha = 0.05$
  - $1 - \alpha/2 = 0.975$
  - We want to find $z_{0.975}$

```
qnorm(0.975)
```

```
## [1] 1.96
```

- How do we find the multiplier for a 90% interval?
  - $\alpha = 0.10$
  - $1 - \alpha/2 = 0.95$
  - We want to find $z_{0.95}$

```
qnorm(0.95)
```

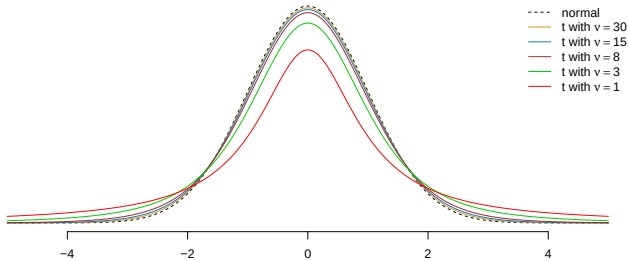
```
## [1] 1.64
```

GAG concentrations

- Let's find an interval estimate for mean GAG concentration!
- We can't... we don't know σ
  - Population standard deviation
- Can we just replace σ with s?
  - No, the sampling distribution is no longer normal
    - All is not lost: most of the reasoning we worked through remains the same
- Replacing σ by s introduces additional noise (variability)
  - Sampling distribution no longer normally distributed
  - We need to use a t-distribution instead

t-distribution

- A t-distribution looks a lot like a (standard) normal distribution
  - Has fatter tails
- Additional parameter $\nu > 0$, called the degrees of freedom
  - This defines how fat the tails are



Historical excursion: William Gosset (1876 – 1937)

- Head Brewer of Guinness who 'discovered' the t-distribution
- Running experiments on yield of barley varieties and did not have statistical tools he needed to analyze the data
  - Statistical methodology developed due to applications in food science, agriculture
- The t-distribution is commonly known as Student's t-distribution
  - Gosset published under the pseudonym 'Student'
  - Guinness allowed its scientists to publish research if they did not mention:
    - Beer
    - Guinness
    - Their own surname

Confidence interval: unkonwn σ

- Replacing σ by s leads to the confidence interval

$$\bar{y} \pm t_{\nu,1-\alpha/2} \frac{s}{\sqrt{n}}$$

- $t_{\nu,1-\alpha/2}$: multiplier for the t-distribution
  - Significance level α
  - Degrees of freedom ν
- When finding confidence interval for μ
  - Degrees of freedom $\nu = n - 1$
- Find multiplier in R: for 95% interval when $n = 30$

```
n = 30
qt(0.975, df = n-1)
```

```
## [1] 2.05
```

GAG concentrations

- We are now ready to find an interval estimate for mean GAG concentration
- We need to get a few bits and pieces together:
  - Call in the data:

```
lect4GAG = read.csv('lect4GAG.csv')
```

  - Find the sample mean: $\bar{y}$

```
ybar = mean(lect4GAG$conc)
ybar
## [1] 2.36
```

  - Find the sample standard deviation: s

```
s = sd(lect4GAG$conc)
s
## [1] 0.668
```

GAG concentrations

- Find the sample size: n

```
n = length(lect4GAG$conc) # length() tells us the number of values
n
## [1] 314
```

- Find the standard error: $s_{\bar{y}} = \frac{s}{\sqrt{n}}$

```
se = s/sqrt(n)
se
## [1] 0.0377
```

- Find the multiplier: 95% confidence interval

```
alpha = 0.05
tcrit = qt(1-alpha/2, df = n-1)
tcrit
## [1] 1.97
```

GAG concentrations

- Put it all together

```
lower = ybar - tcrit * se # lower confidence limit
upper = ybar + tcrit * se # upper confidence limit
ci = c(lower, upper)
ci
## [1] 2.29 2.44
```

- The 95% confidence interval for μ is (2.29, 2.44)
  - Interval estimate for μ

- Spent some time interpreting the interval in the next lecture

Summary

- Found confidence interval for μ
  - Interval that quantifies how precise our estimate of μ is
- Found confidence interval if σ is known
  - Useful for understanding
  - Not practically useful
- Found confidence interval if σ is unknown
  - Introduced the t-distribution
- Looking forward:
  - More about confidence intervals