STAT115: Introduction to Biostatistics

University of Otago

Ōtākou Whakaihu Waka

Lecture 11: The Normal Distribution

Outline

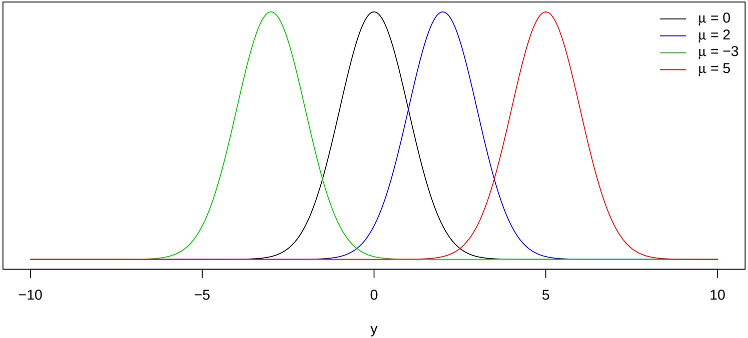
- Previous lectures:
  - Introduction to probability, random variables
  - First example of a statistical model
    - Normal model
- Today: learn more about the normal distribution

Normal distribution

- We used a normal model to describe total cholesterol in male heart attack patients
- 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

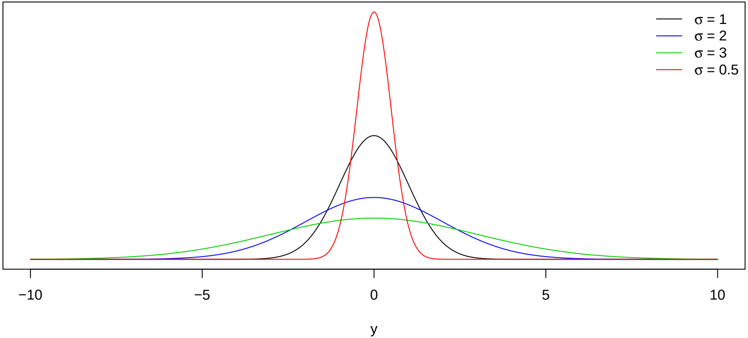
Recap: normal distribution

- 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

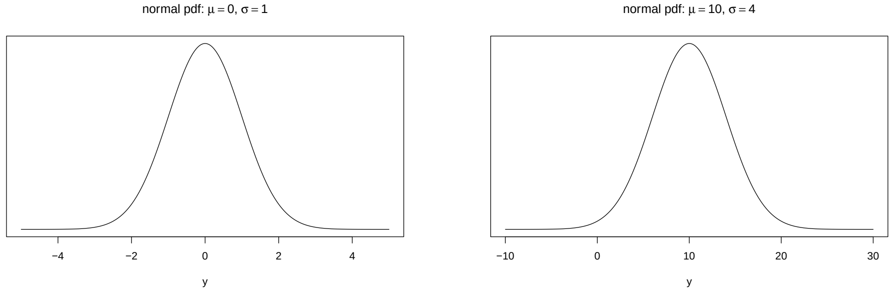
Probabilities

- 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

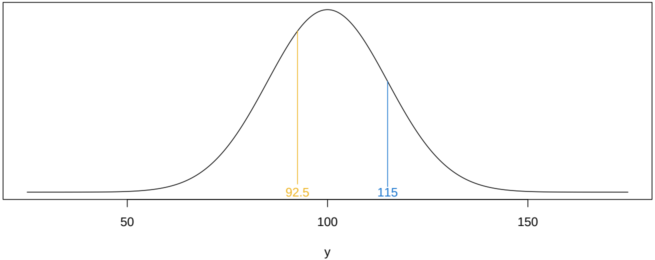
$$y = \mu + z\sigma$$

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

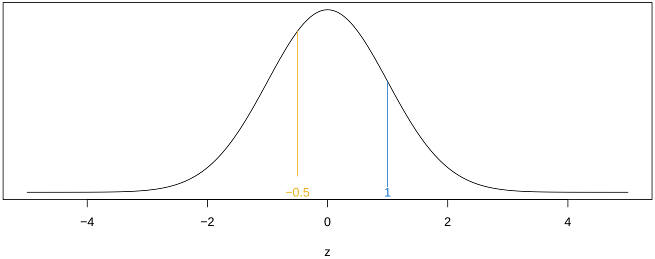
$$y = 100 + 1 \times 15 = 115$$

Graphical representation

normal pdf: $\mu = 100$, $\sigma = 15$

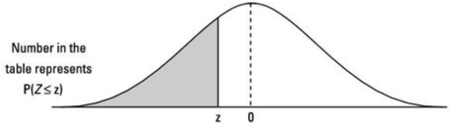


normal pdf: $\mu = 0$, $\sigma = 1$



Finding probabilities: deep dark past

- We used to find probabilities from tables



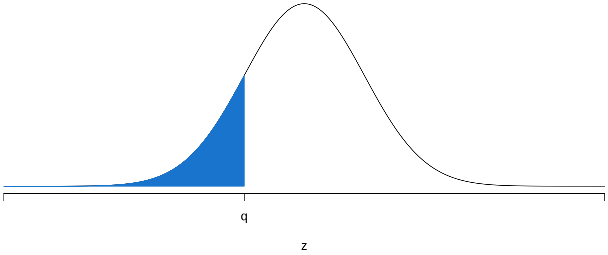
z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
–3.6	.0002	.0002	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001
–3.5	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002
–3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
–3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.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 115, 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:

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

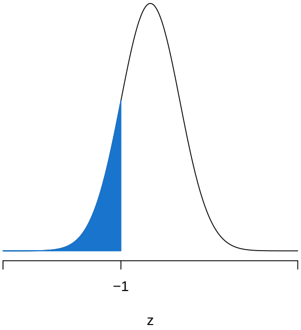
- Find $\Pr(Z < -1)$

```
mu = 100; sigma = 15 # the mean and sd for IQ
z = (85 - mu)/sigma # finding the z-score
pnorm(z)

## [1] 0.1587

pnorm(-1) # for those who want to check

## [1] 0.1587
```



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)

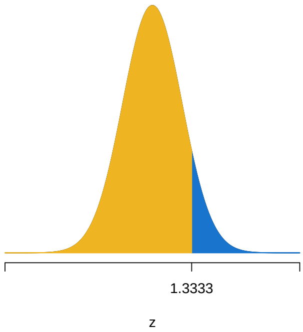
```
z = (120 - mu)/sigma # finding the z-score
pnorm(z)

## [1] 0.9088
```

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

```
1-pnorm(z)

## [1] 0.09121
```



Example 3

- Probability that IQ is between 110 and 130?

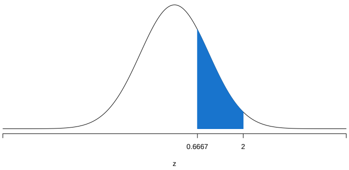
$$z_{110} = \frac{y-\mu}{\sigma} = \frac{110-100}{15} = 0.6667$$
$$z_{130} = \frac{y-\mu}{\sigma} = \frac{130-100}{15} = 2$$

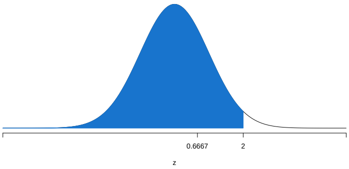
```r
z110 = (110 - mu)/sigma # finding the z-score
z130 = (130 - mu)/sigma
```

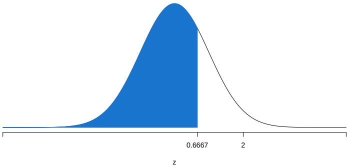
- $\Pr(z_{110} < Z < z_{130}) = \Pr(Z < z_{130}) - \Pr(Z < z_{110})$
  - Best seen graphically on RHS

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

```
## [1] 0.2297
```





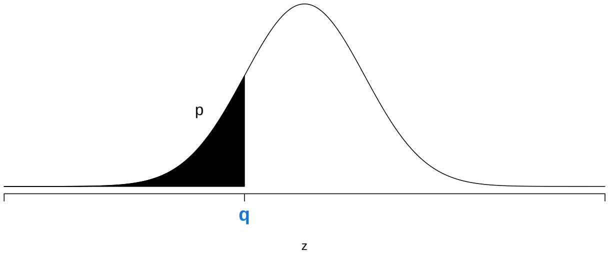


Important properties

- We can use this to learn some important characteristics of a normal distribution
- $\Pr(-1 < Z < 1) = 0.6827$
  - Approximately 68% of values should be within 1 sd of the mean
- $\Pr(-2 < Z < 2) = 0.9545$
  - Approximately 95% of values should be within 2 sd of the mean
- $\Pr(-3 < Z < 3) = 0.9973$
  - 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 $p = 0.98$

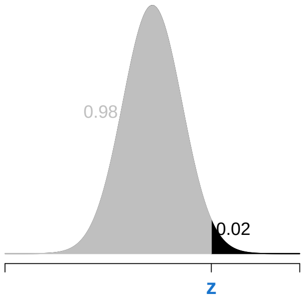
```
z = qnorm(0.98)
```

- Find the y-value

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

```
mu + z * sigma
## [1] 130.8
```

- 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] 0.1587
```

- 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 $\bar{y}$?