Confidence Intervals

Statement for Audio and Video Learning Resources

Video and audio content at the University uses closed captions generated by automatic speech recognition (ASR). The ASR process is based on machine learning algorithms which automatically transcribe voice to text. According to our technology providers, this process is approximately 70-90% accurate depending on the quality of the audio, and consequently video and audio closed captions may include some transcription errors. It is therefore important to recognise that the original recording is the most accurate reflection of the content, and not the captions.

If you require accurate captions as part of your reasonable adjustments, please contact the Inclusion Centre to discuss your requirements.

Content

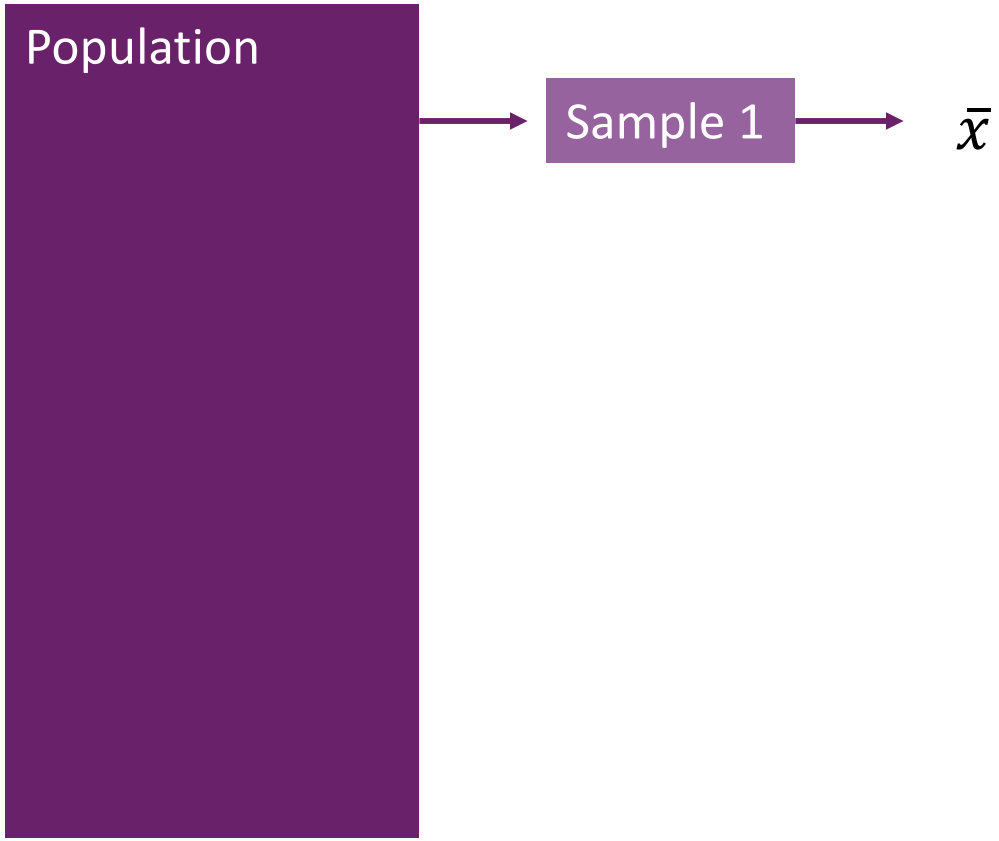
- Populations vs Samples
- Hypothesis testing
  - Definitions
  - Confidence intervals
  - [more on hypothesis testing on week 9]

Populations vs Samples

- Generally not all data is available
- Observations in data
  - Are a subset (sample set) of all observations (population)
  - If this sample set is representative of the population, estimations about the population can be drawn.

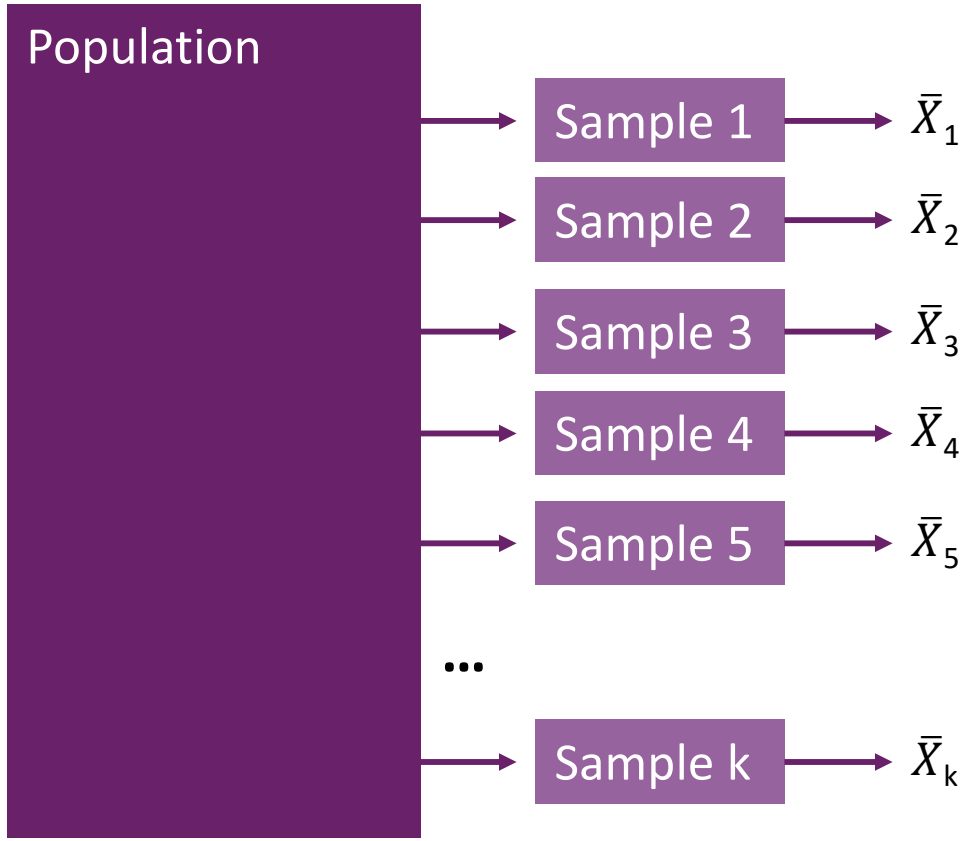
Samples

- Given a population
  - Select ONE sample
  - Calculate mean $\bar{x}$
  - How close is $\bar{x}$ to the population mean μ?



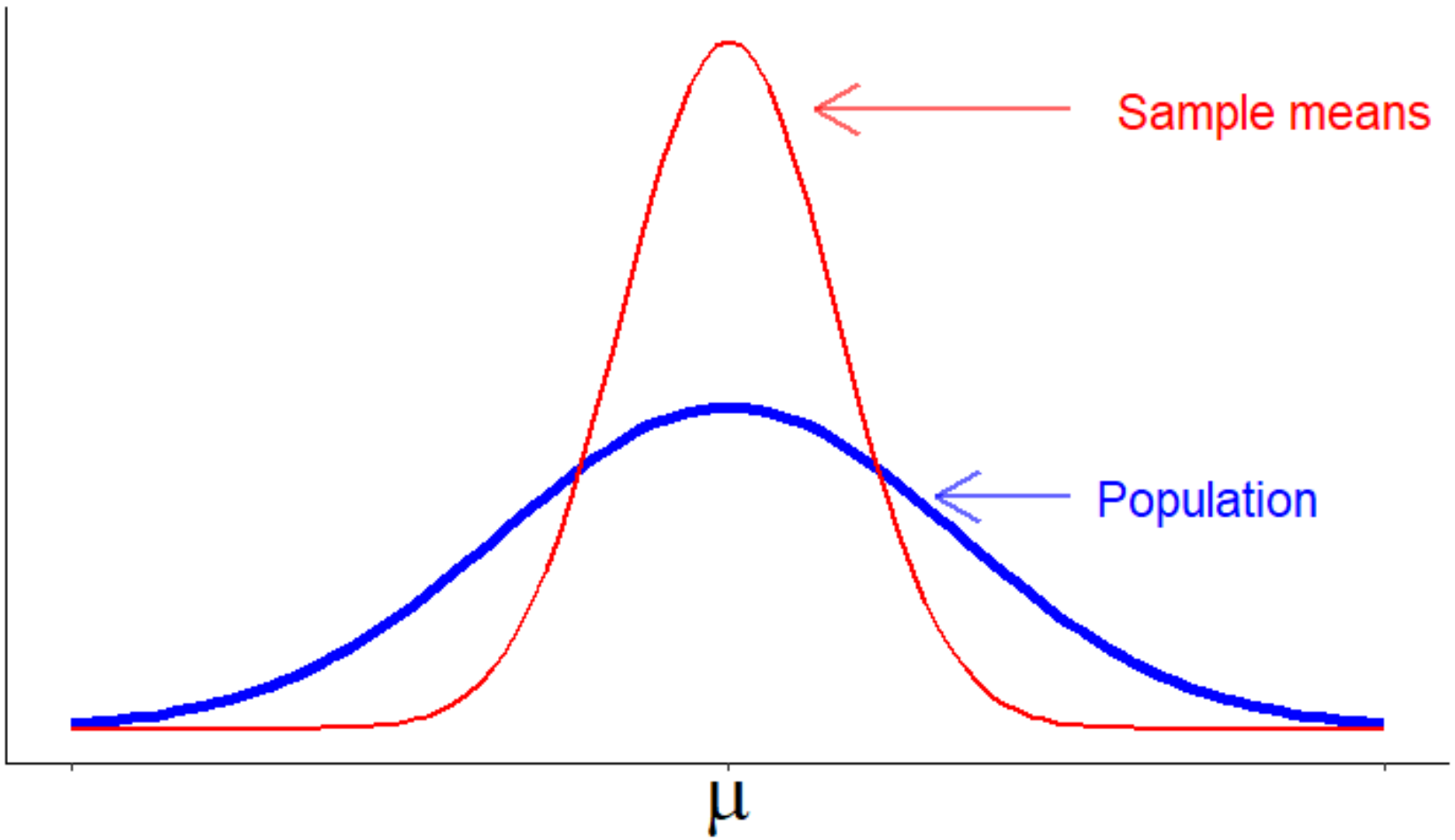
… samples

- Select several samples
- Calculate mean from each
- Have set of means of samples
  - Means unlikely to be identical
- Set of sample means
  $\{\bar{X}_1, \bar{X}_2, \bar{X}_3, \bar{X}_4, \bar{X}_5 \ldots \bar{X}_k\}$
- Mean of sample means distribution $\mu_{\bar{x}}$
- Standard deviation of sample means distribution $\sigma_{\bar{X}}$



Standard error of the mean

- Standard deviation of the distribution of sample means < standard deviation of individual samples
  - Means of samples closer to "middle" than individual items.
- $\sigma_{\bar{\bar{X}}} = \frac{\sigma}{\sqrt{n}}$
- Estimated as $\frac{s}{\sqrt{n}}$



Population vs sample means

Population

- Individual items X
- Mean μ
- Standard deviation σ
- Points less close to each other

Sample means

- Sample means items $\bar{X}$
- Mean $\mu_{\bar{x}} = \mu$
- Standard deviation $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$
- Points are closer together

Standard deviation

- For whole population
  - $\sigma_n = \sqrt{\frac{\sum(x - \bar{x})^2}{n}}$
- Sample standard deviation
  - $\boldsymbol{s} = \sigma_{n\_1} = \sqrt{\frac{\sum(x - \bar{x})^2}{\boldsymbol{n-1}}}$

Single sample

- In practice, single sample is selected from population
- Sample mean provides estimate of population mean
- But population mean unlikely to be exactly the sample mean
  - $\mu = \bar{X} \pm$ margin or error

- **Confidence interval for a population mean:** a range of values within which the population mean lies with a specified level of confidence.
  - Higher confidence → wider interval
  - Lower confidence → smaller interval
- Confidence level: 90%, **95%** and 99% are frequently used.

How good is a conclusion?

- Assume that a dataset contains 20 entries about staff at RGUOils – a company with 1000 staff.
- From the dataset you calculate that on average, staff travel 3.2 miles to work.
  - Is 3.2 miles representative of the distance travelled by staff?
  - How close is this average to the real average (the one for all staff)?

Confidence

- In reality, the real value is unlikely to be exactly 3.2 miles, but it will be a little more or a little less.
- As only 20 observations are collected, the real average may not be very close to 3.2
- If there were 100 observations, our confidence in the result being close to 3.2 would be bigger.

… confidence intervals

- Measure uncertainty
  - Describe uncertainty in estimates
- What is 95% confidence?
  - If multiple samples taken, getting a confidence interval for each
  - 95% of the confidence intervals would contain the true sample.

Confidence interval

- An interval which is expected to contain the true value.
  - E.g. (0.5, 5.9) for average travel distance
  - The true average distance travelled by staff is around 3, and is contained in interval (0.5, 5.9) with some degree of confidence.

Example

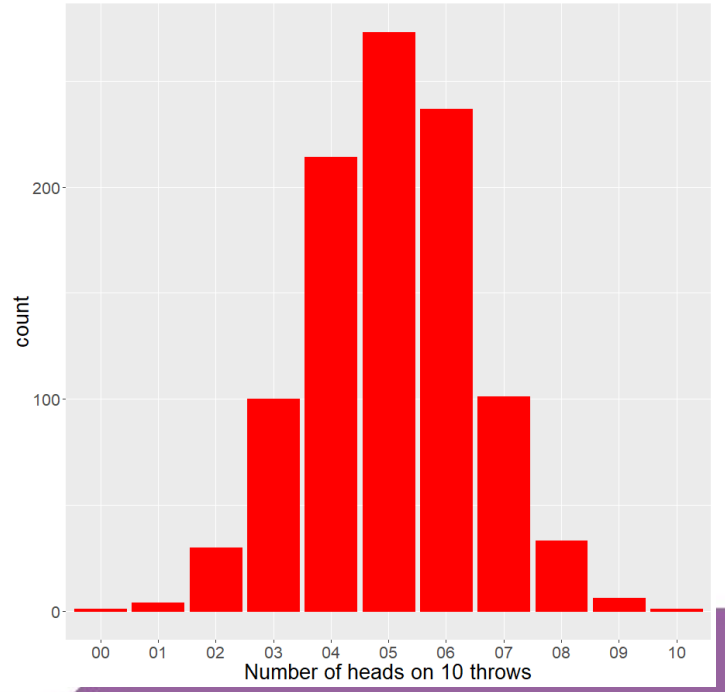
- Coin tossing experiment
  - Toss coin 10 times
  - Repeat experiment 100 times
- If coin is not biased, expect 50% heads
  - But each run will have a different value
  - On average, ~50% will be heads, i.e. 5 heads

Results of coin tossing 10 times – repeated 1000 times

- 4/10
- 5/10
- 3/10
- 6/10
- 5/10
- 7/10
- 5/10
- 6/10
- 5/10
- 4/10
- … [up to 1000 times]

4 heads out of 10 tosses

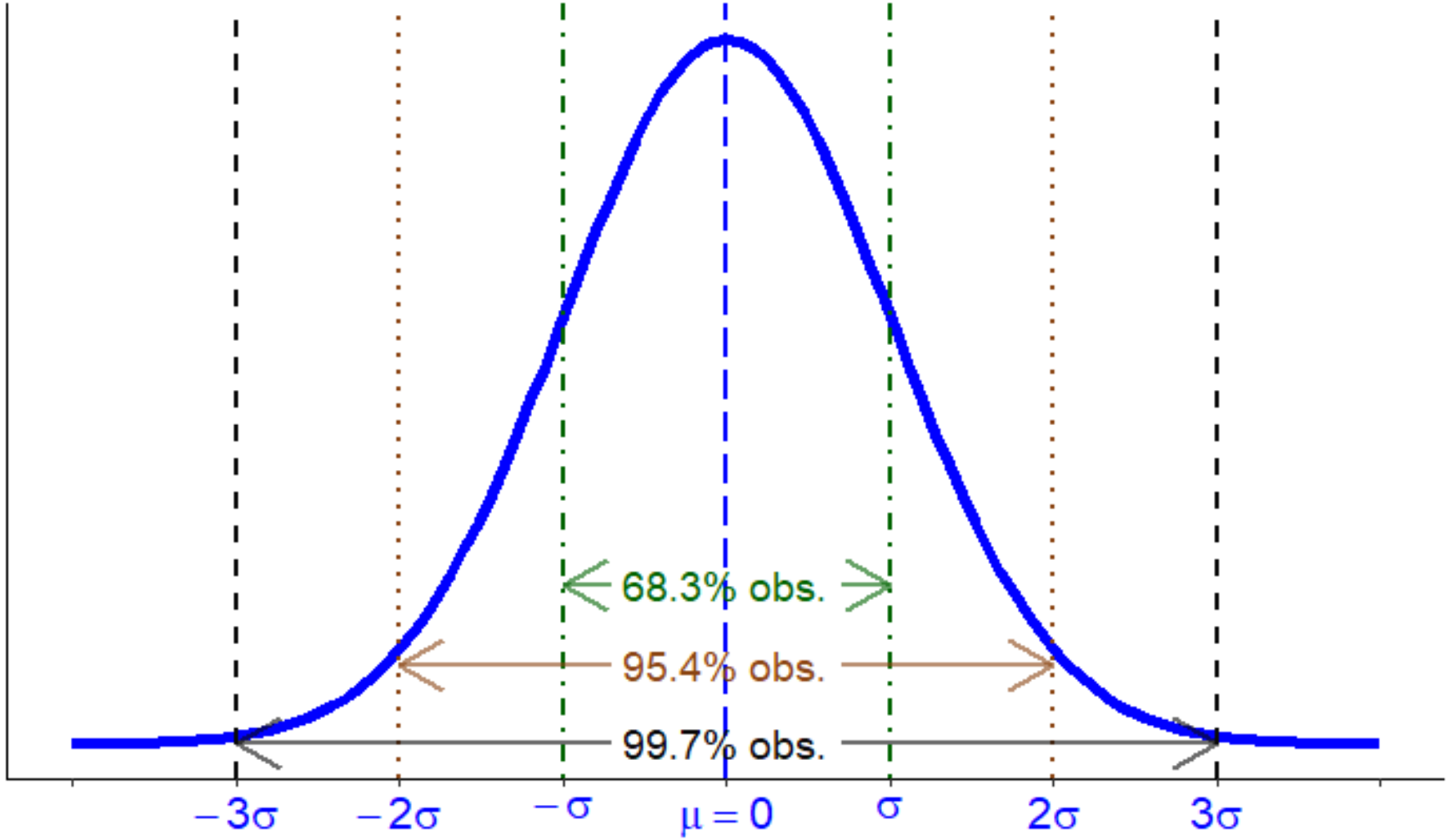
- Mean (assume 1000 repetitions or samples)
- $\frac{4+5+3+6+5+7+5+6+5+4+\ldots}{1000} \simeq 5$
- The more repetitions (samples) the more likely the average is 5
- "~Normal" distribution
- Confidence interval at
  - 95%
    - **Out of 1000 samples if a confidence interval is calculated for each sample, 95% of them would contain the true value.**



Sampling

- Sampling distributions – if sample is large generally
  - Distribution resembles normal distribution
  - Calculated standard error (see below) is very accurate
- What if sample size is small?
  - If distribution is nearly normal use ***t-distribution***
    - Enables use of statistics to measure uncertainty
    - ***df – degrees of freedom*** – number of values that are free to change.
    - If n = size of sample, df= n-1
    - The higher the value of n, the less uncertainty

Normal distribution



ts of physical measures follow
rmal distribution.

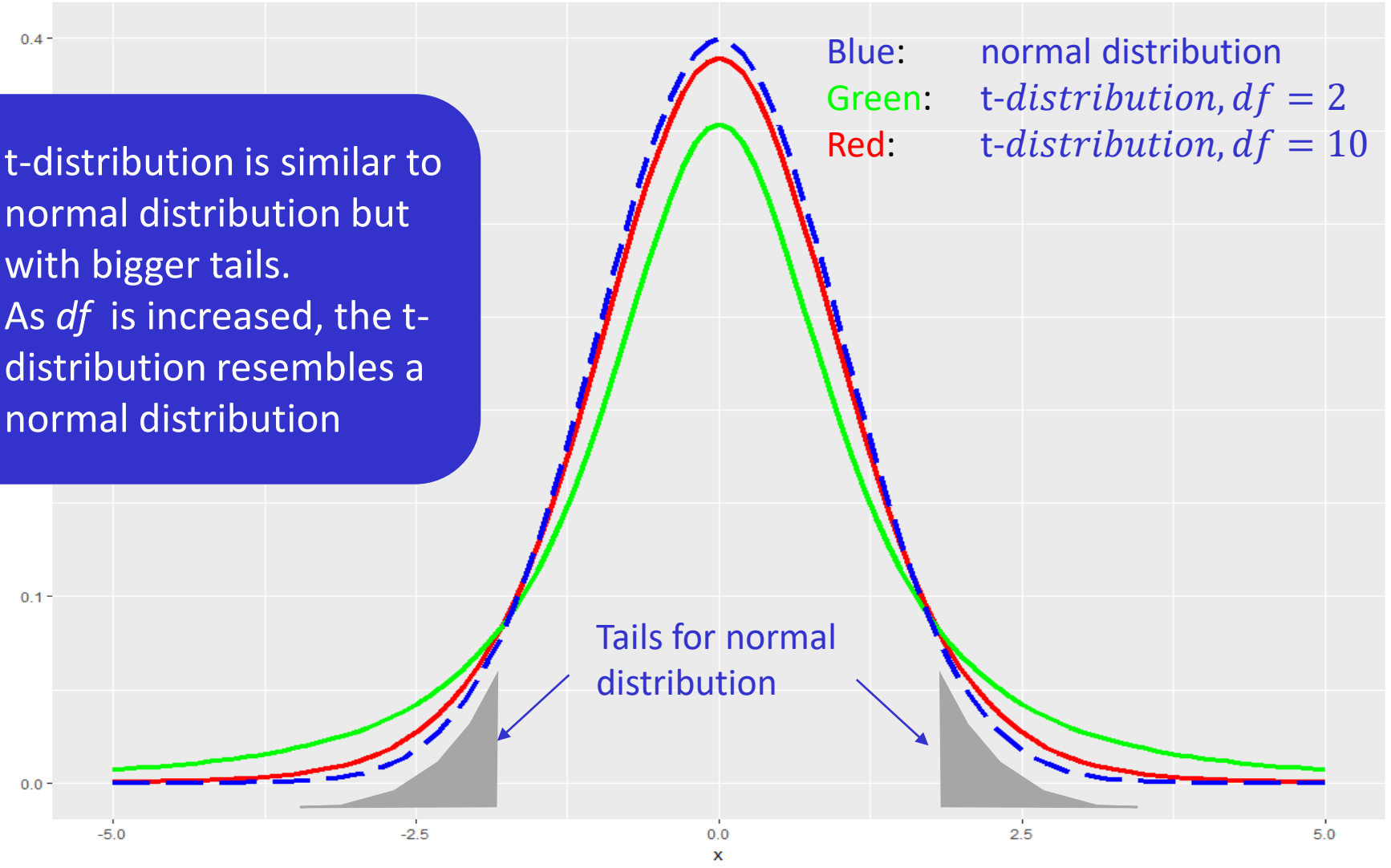
ariable of interest
nean
standard deviation
} often unknown

erences about populations based
samples are often made using

sample means
sample standard deviation.

Normal vs t-distribution

- t-distribution is similar to normal distribution but with bigger tails.
- As *df* is increased, the t-distribution resembles a normal distribution



Confidence interval - formula

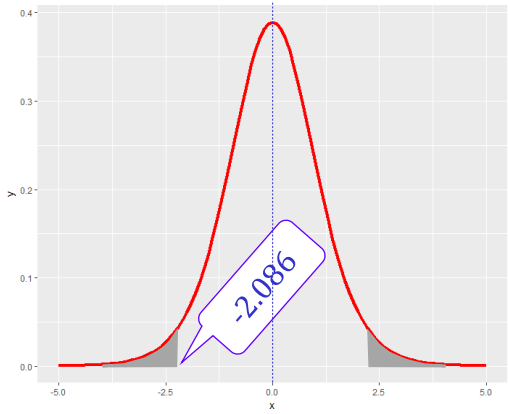
- Given by $\mu = \bar{X} \pm t \frac{s}{\sqrt{n}}$
  - Where
  - $\bar{X}$ *is the sample mean*
  - s is the sample standard deviation
  - n is the sample size
  - t is a value from the t-distribution table with v=n-1 degrees of freedom
    - Look-up value in table
    - Calculate significance level α = (1 - confidence level)
    - E.g. 95% confidence, α = 1 - 0.95 = 0.05
    - Look at two-tail corresponding value for the required degrees of freedom
    - Why 2 tail? Half the interval is below $\bar{X}$, half the interval is above $\bar{X}$

t-table

DF	One tail	0.1	0.05	0.025	0.01	0.005	0.001	0.00025
	Two tail	0.2	0.10	0.05	0.02	0.01	0.002	0.001
1		3.078	6.314	12.706	31.821	63.656	318.289	636.578
2		1.886	2.920	4.303	6.965	9.925	22.328	31.600
3		1.638	2.353	3.182	4.541	5.841	10.214	12.924
4		1.533	2.132	2.776	3.747	4.604	7.173	8.610
5		1.476	2.015	2.571	3.365	4.032	5.894	6.869
6		1.440	1.943	2.447	3.143	3.707	5.208	5.959
7		1.415	1.895	2.365	2.998	3.499	4.785	5.408
8		1.397	1.860	2.306	2.896	3.355	4.501	5.041
9		1.383	1.833	2.262	2.821	3.250	4.297	4.781
10		1.372	1.812	2.228	2.764	3.169	4.144	4.587
11		1.363	1.796	2.201	2.718	3.106	4.025	4.437
12		1.356	1.782	2.179	2.681	3.055	3.930	4.318
13		1.350	1.771	2.160	2.650	3.012	3.852	4.221
14		1.345	1.761	2.145	2.624	2.977	3.787	4.140
15		1.341	1.753	2.131	2.602	2.947	3.733	4.073
16		1.337	1.746	2.120	2.583	2.921	3.686	4.015
17		1.333	1.740	2.110	2.567	2.898	3.646	3.965
18		1.330	1.734	2.101	2.552	2.878	3.610	3.922
19		1.328	1.729	2.093	2.539	2.861	3.579	3.883
20		1.325	1.725	**2.086**	2.528	2.845	3.552	3.850
...								

t-table interpretation - example

- Assume t-distribution (2 tailed) with df=20.
- The cut-off for a 5% (0.05)
  - Upper tail is **2.086**
  - Lower tail is **-2.086**
  - 95% is within the interval
  - So overall 5% outside the interval, 2.5% below, 2.5% above
  - 95% confidence



Central limit theorem for normal data

- *"The sampling distribution of the mean is nearly normal when the **sample observations** are **independent** and come **from a nearly normal distribution**. This is true **for any sample size**."*

Conditions

- Independence of observations
  - Sample contains **< 10% of population** obtained at random
- Check distribution is nearly normal
  - Plot data, check skews
  - Consider previous data (experience with the data)
  - If sample contains n observations use $\mathbf{df = n\text{-}1}$
- Relax conditions for sample size >= 30
  - A little skew may be OK
  - If obvious outliers, n should be >= 100

… confidence intervals (cont.)

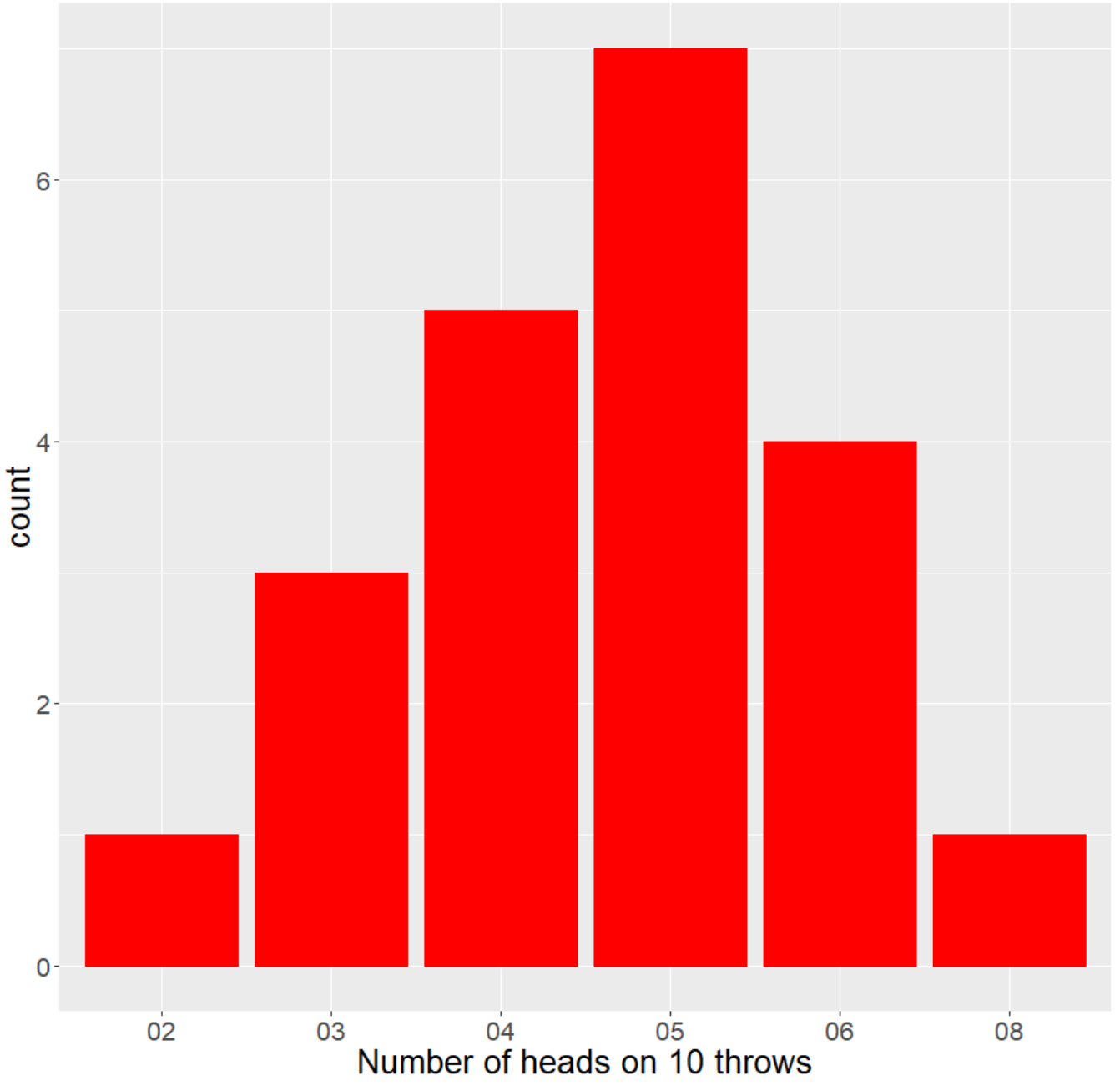
- Calculate mean $\bar{x}$
- Establish df
  - $\mathrm{df} = \mathrm{n} - 1$
- **Significance level α** is 1 - confidence level
- Check table for value t_{df} associated with significance level for corresponding df
- Calculate **standard error**
  - $SE = \frac{s}{\sqrt{n}}$
- **Margin of error** is
  - $t_{df} * SE$
- Interval is sample mean ± margin of error
  - $\bar{x} \pm (t_{df} * SE)$

Example of CI

- Assume dataset with following 21 observations regarding coin tossing 10 times
  - 5 5 5 2 5 5 3 4 4 4 5 4 8 6 6 3 5 4 3 6 6
  - Calculate confidence interval at 95% confidence?

Can we apply t-test?

- Coin tossing were independent
- 21 observations is < 10% overall
- Observations are within 2.5 standard deviations of mean (4.6667)
  - $s = \sqrt{\frac{\sum(x - \bar{x})^2}{n-1}} = 1.546006$
- So yes t-test is suitable



Estimating confidence interval

- $\bar{x} = \frac{5+5+5+2+\ldots+3+6+6}{21} = 4.666667$
- $df = 21 - 1 = 20$
- $\alpha = 1 - 0.95 = 0.05$
- $t_{df} = 2.086$ *(from table, find value for two-tailed with* $\alpha = 0.05$*)*
- $s = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}} = 1.354006$
- $SE = \frac{s}{\sqrt{n}} = 0.2954684$
- $\bar{x} \pm (t_{df} * SE) = 4.666667 \pm (2.086 * 0.2954684)$
  $\rightarrow$ ***(4.05032, 5.283014)***
- *At 95% confidence the true mean* $\in$ ***(4.05032, 5.283014)***

What if confidence level is 0.99?

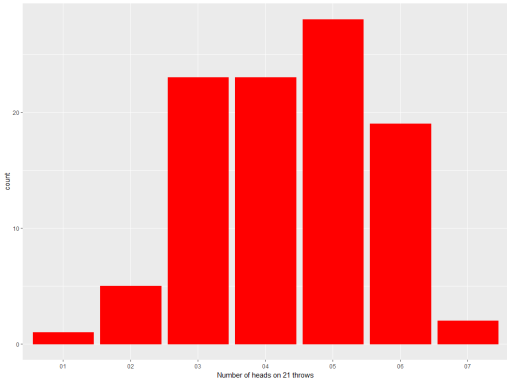
- For 99% confidence
  - $t_{df} = 2.845$
  - $\bar{x} \pm (t_{df} * SE) = 4.666667 \pm (2.845 * 0.2954684)$
    → *(3.825959, 5.507375)*
- For 90% confidence
  - $t_{df} = 1.725$
  - $\bar{x} \pm (t_{df} * SE) = 4.666667 \pm (1.725 * 0.2954684)$
    → *(4.157067, 5.176266)*
- ***Higher confidence → bigger interval***

Result interpretation

- **90% confidence interval** – if sampling the population 100 times as above, on average 90% of the time the interval ***(4.157067, 5.176266)*** will include the true population mean.
- **95% confidence interval** – if sampling the population 100 times as above, on average 95% of the time the interval ***(4.05032, 5.283014)*** will include the true population mean.
- **99% confidence interval** – if sampling the population 100 times as above, on average 99% of the time the interval ***(3.825959, 5.507375)*** will include the true population mean.

Example (biased?)

- Assume dataset with 101 observations regarding coin tossing 10 times
  - Is coin biased at 90% confidence?
  - $\bar{x} = 4.356436$ $min = 1, max = 7$
  - $df = 100$
  - $t_{100} = 1.290$
  - $s = 1.269521$
- Does plot show a normal distribution?
  - Shape not entirely normal
  - Valid test?
- Values are within 2.7 standard deviations from mean



Estimating confidence interval example

- $\alpha = 1 - 0.90 = 0.1$
- $SE = \frac{s}{\sqrt{n}} = 0.126322$
- $\bar{x} \pm (t_{df} * SE) = 4.356436 \pm (1.290 * 0.126322)$

  $\rightarrow$ ***(4.19348062, 4.51939138)***

  We are 90% confidence that true mean $\in$ ***(4.19348062, 4.51939138)***

  Different from (never associate to a probability!)

  ~~*There is a 90% probability that true mean* $\in$ *(4.19348062, 4.51939138)*~~

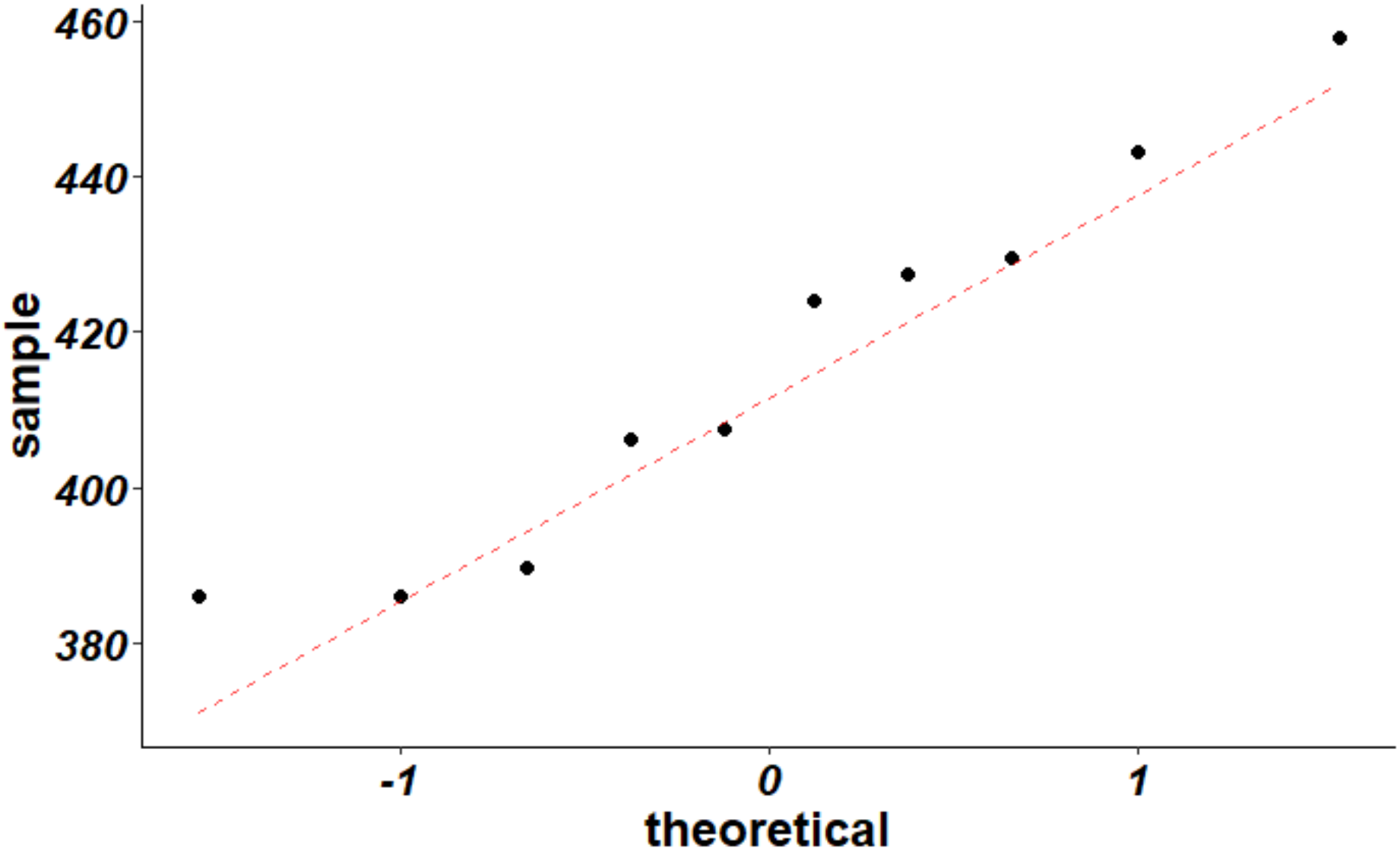
  Interval does not contain 5 so coin likely to be biased.

Constraints

- When n < 30, intervals are valid if the data are randomly sampled from a population with a normal distribution.
- Checks
  - Distribution of values – is it bell-shaped?
  - Q-Q plot – sample against theoretical normal distribution. Points should be close to straight line.
  - Shapiro-Wilk normality test
    - If p-value is greater than significance level, it is reasonable to assume that the distribution is normal – see lab.

QQ-Plot

- Sample against theoretical normal distribution.
- Points close to a straight line.
  - Normal distribution



Summary

- Confidence intervals – show mean and uncertainty
- A confidence interval at 95% indicates that if multiple samples were taken, getting a confidence interval for each
  - 95% of the confidence intervals would contain the true sample.
- Confidence is NOT the same as probability.