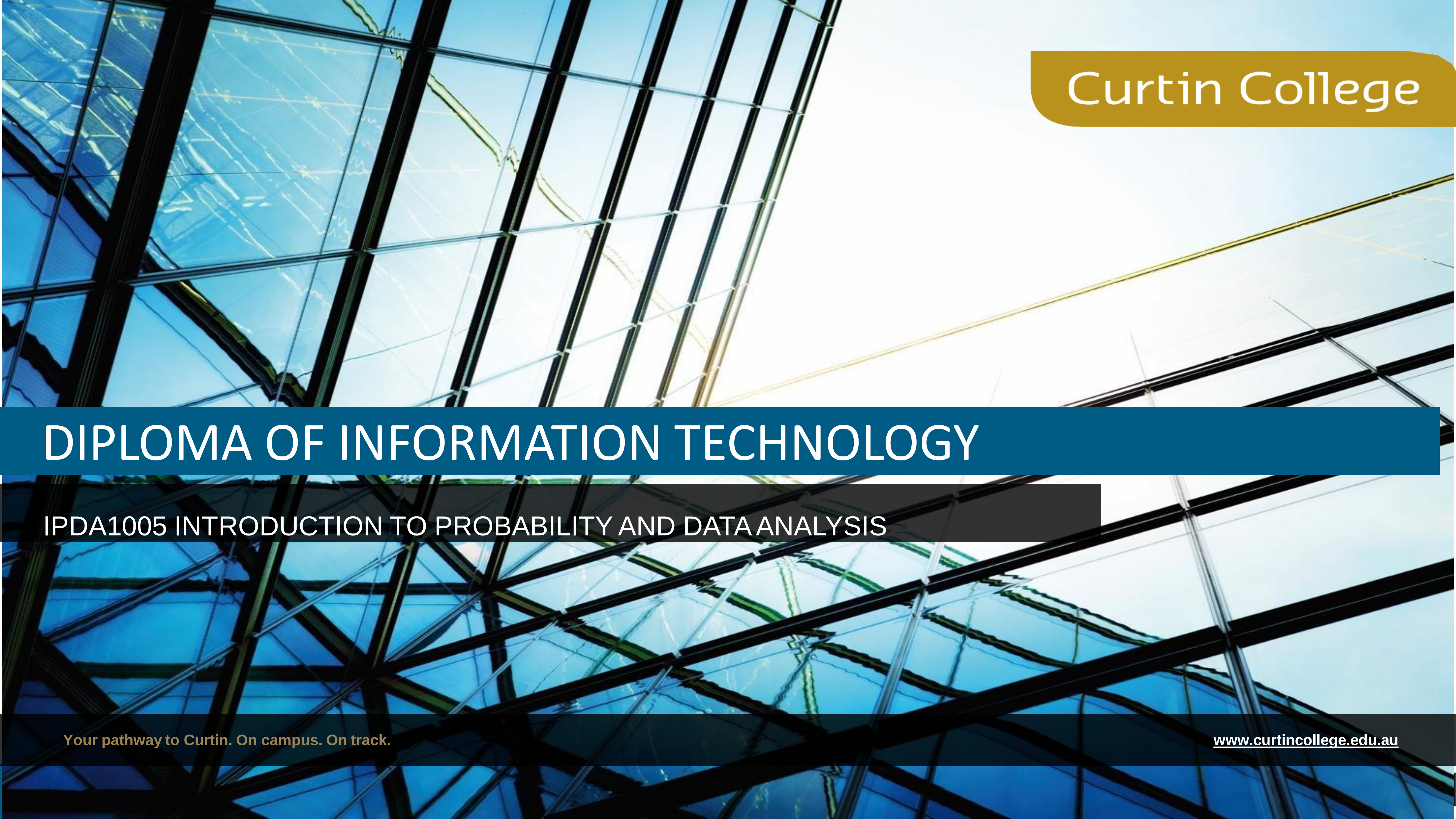


Curtin College

DIPLOMA OF INFORMATION TECHNOLOGY

IPDA1005 INTRODUCTION TO PROBABILITY AND DATA ANALYSIS

Acknowledgement

We respectfully acknowledge the Elders and custodians of the Whadjuk Nyungar nation, past and present, their descendants and kin. Curtin College Bentley Campus enjoys the privilege of being located in Whadjuk / Nyungar Boodjar (country) on the site where the Derbal Yerrigan (Swan River) and the Djarlgarra (Canning River) meet. The area is of great cultural significance and sustains the life and well being of the traditional custodians past and present.

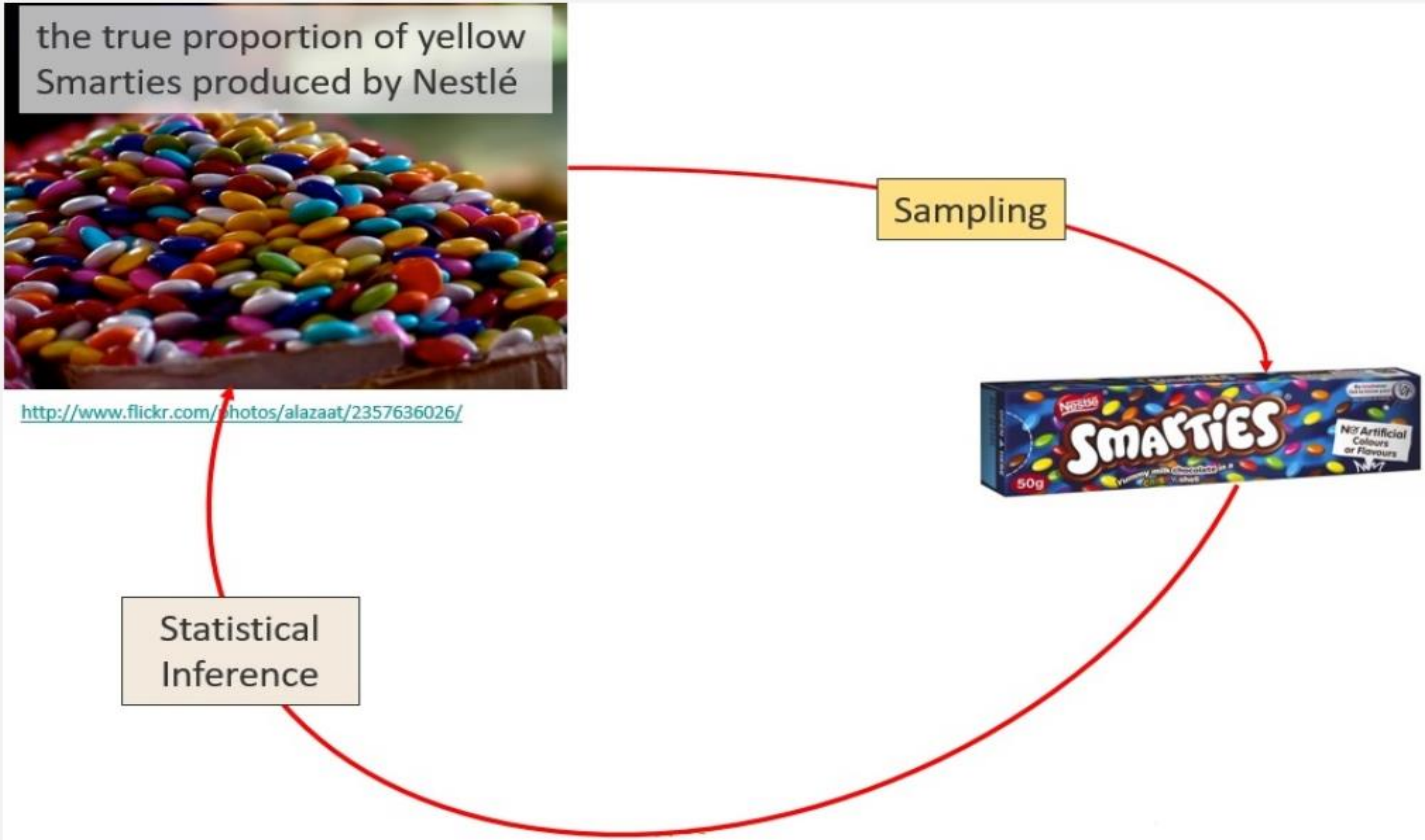
Outline

1. Population and Sample
2. Parameters and Statistics
3. Sampling distributions
4. Central Limit Theorem
5. Confidence Interval for a Proportion
6. Choosing the Sample Size
7. Confidence Intervals for Mean

Population and Sample

- Suppose we want to estimate the proportion of yellow Smarties among all Smarties produced by Nestlé.
- Here the entire collection of Smarties produced by Nestlé is considered to be the *population* and, in this context, the proportion of yellow Smarties is the *population proportion* of interest.
- We can select a random *sample* from this population, and find the proportion of yellow Smarties in this sample. This is the *sample proportion*.
- The sample proportion might be used as an estimate of the population proportion.
- Assuming that Smarties packets are filled at random, we could use a packet of Smarties as a random sample. If we need a larger sample we just use more packets.

Proportion of Yellow Smarties



the true proportion of yellow Smarties produced by Nestlé

http://www.flickr.com/photos/alazaat/2357636026/

Sampling

Statistical Inference

Parameters and Statistics

- A *parameter* is a characteristic of the population. In practice, the value of a parameter is -
  - constant
  - unknown (because we cannot examine the entire population)
- A *statistic* is a characteristic of a sample. The value of a statistic is -
  - variable (from sample to sample)
  - known (we calculate it from the sample)
- Examples:
  - Mean blood pressure of Australian adult males
  - Proportion of voters who will vote for independent candidates
- In order to use a statistic for inference about a parameter, we must understand how it varies. That is, we want to know its probability distribution. Because statistics come from samples, their distributions are known as *sampling distributions*.

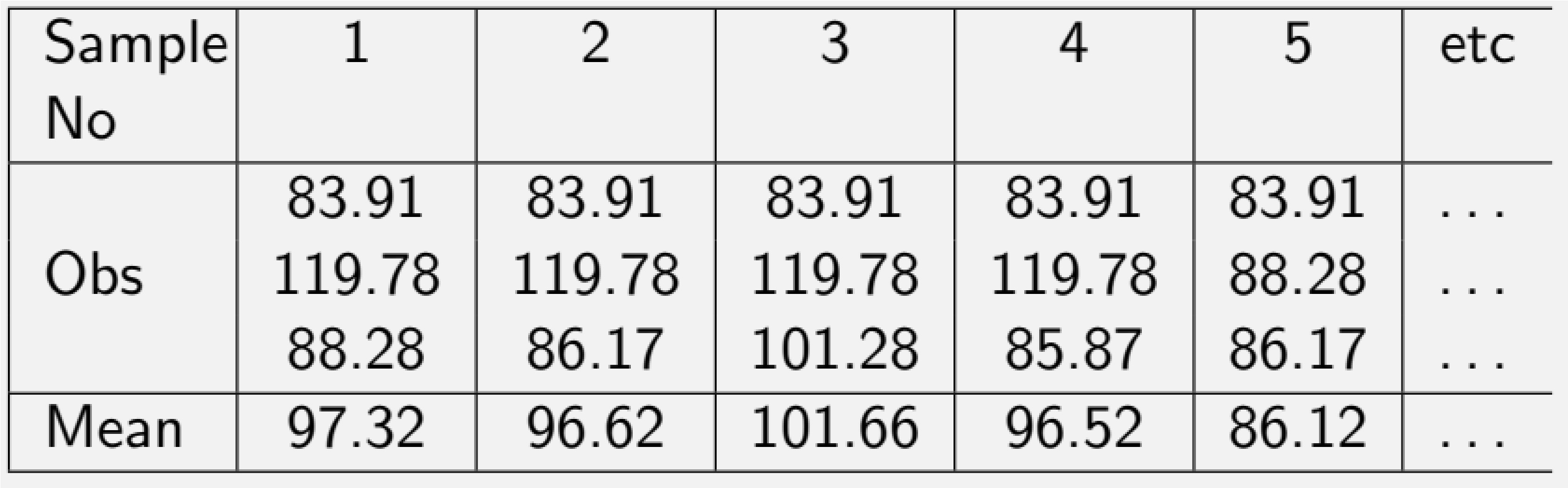
Sampling Distributions

Once we have a sample of data, we know the size of the sample. Then, to use the sample mean (say) for inference about the population mean, we need to describe the sampling distribution of a mean in samples of a certain size.

- Here is a population of 6 measurements:

  83.91 119.78 88.28 86.17 101.28 85.87

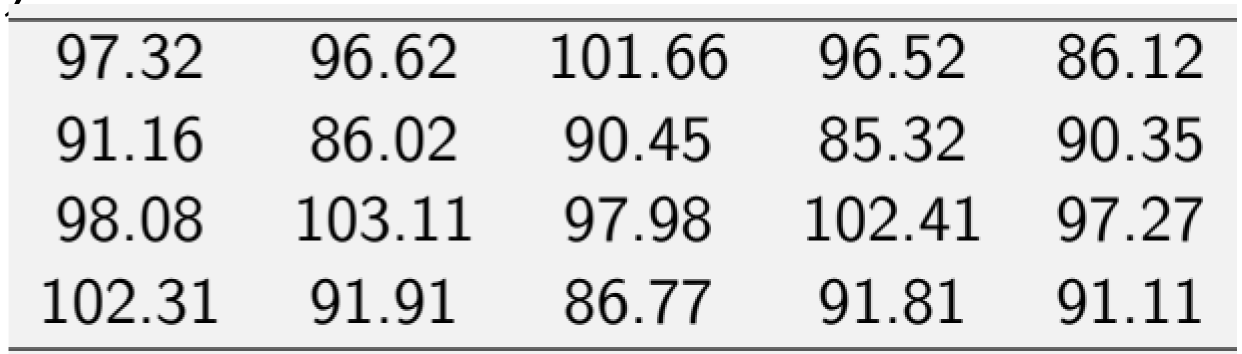
- We select samples of size *n* = *3* from this population, and calculate the mean each time:



Sample No	1	2	3	4	5	etc
Obs	83.91	83.91	83.91	83.91	83.91	. . .
	119.78	119.78	119.78	119.78	88.28	. . .
	88.28	86.17	101.28	85.87	86.17	. . .
Mean	97.32	96.62	101.66	96.52	86.12	. . .

Sampling Distributions

- There are $\binom{6}{3} = 20$such samples, and here are their means:



97.32	96.62	101.66	96.52	86.12
91.16	86.02	90.45	85.32	90.35
98.08	103.11	97.98	102.41	97.27
102.31	91.91	86.77	91.81	91.11

- Since these are the means of all possible samples of size 3 from our population, we can treat them as a population of sample means. For the sample means, we have μ_X= 94.215, $\sigma_X = 5.862$.
- Compare this to the population values

83.91 119.78 88.28 86.17 101.28 85.87

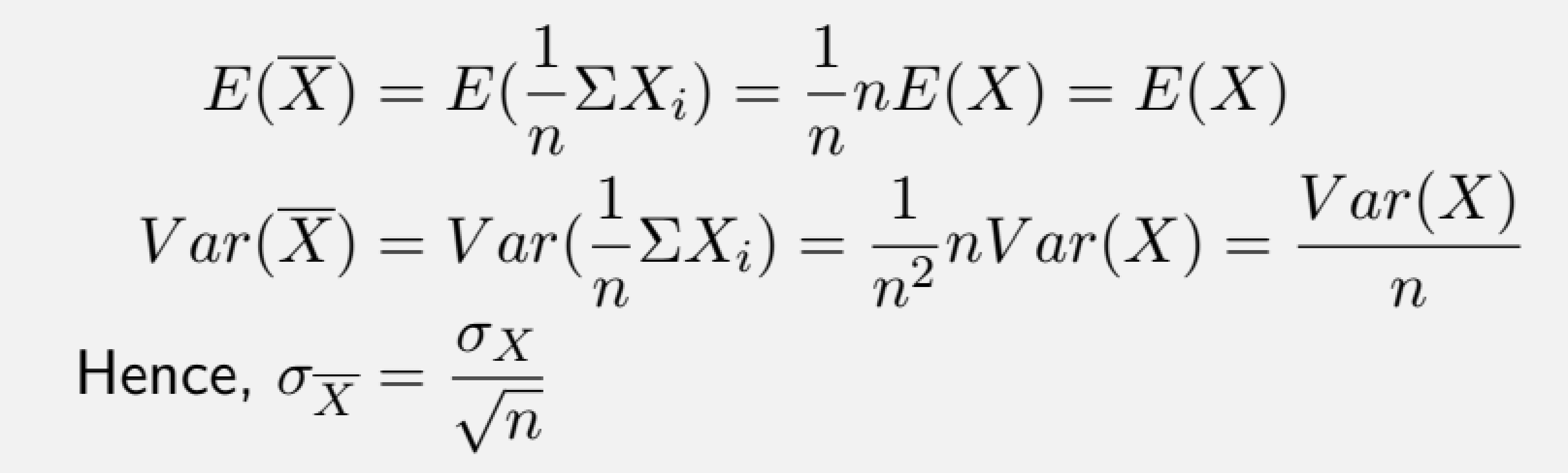
for which we have $\mu_X = 94.215, \sigma_X = 13.995$.

- The mean of the sample means is equal to the population mean $\mu_{\bar{X}} = \mu_X$

The standard deviation of sample means is less than the population standard deviation $(\sigma_{\bar{X}} = \sigma_X)$.

Sampling distribution of mean

- From observing the results of one experiment, I said $\mu_{\bar{X}} = \mu_X$ and $\sigma_{\bar{X}} < \sigma_X$. But we can do better than that.
- $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$and we are obtaining the x_i's by sampling from the same population, so each x_i has the same mean and variance as the population.
- Using earlier results on expectation and variance, and if the x_i's are independent, we can prove that



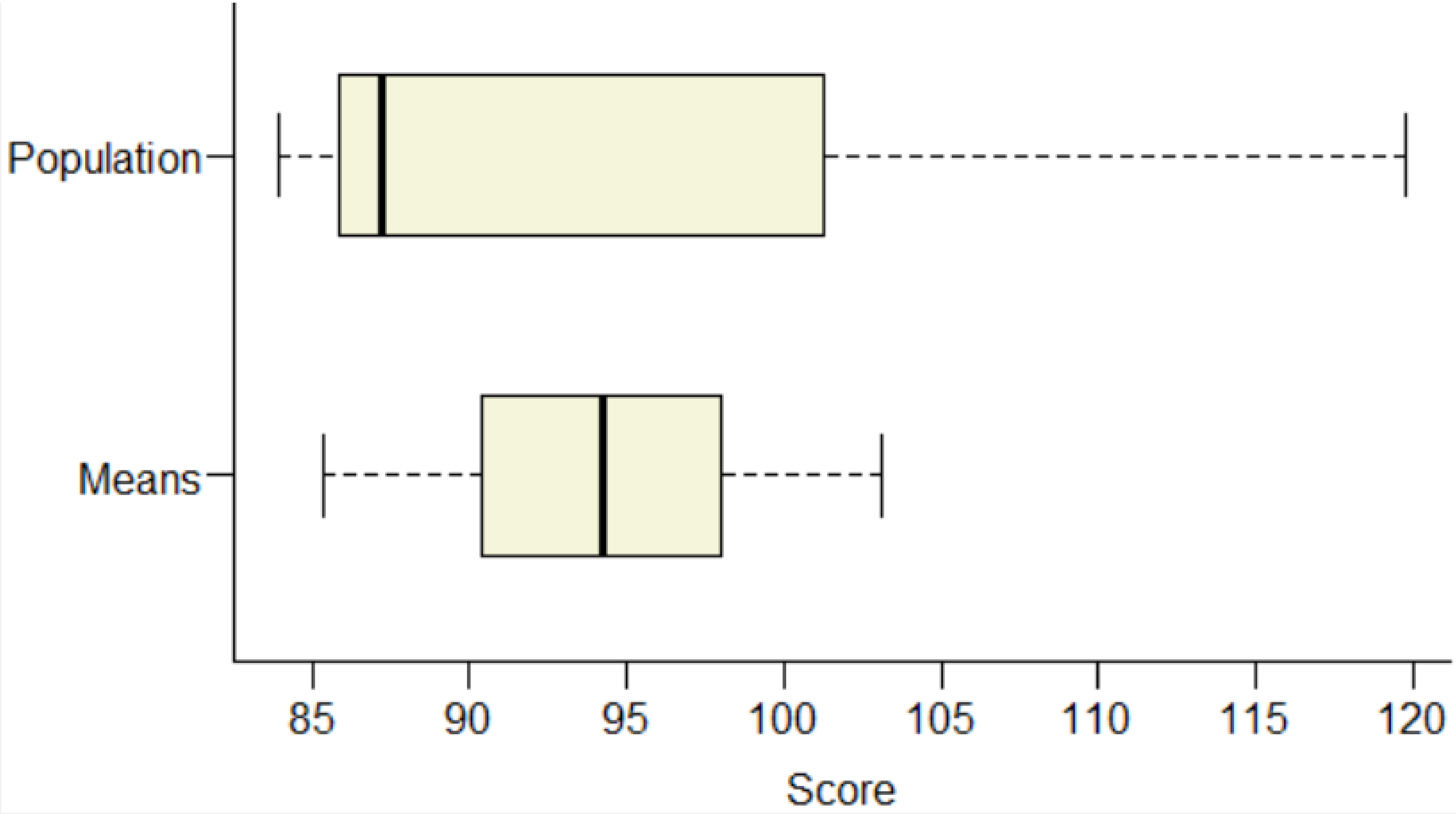
$$E(\overline{X}) = E(\frac{1}{n}\Sigma X_i) = \frac{1}{n}nE(X) = E(X)$$

$$Var(\overline{X}) = Var(\frac{1}{n}\Sigma X_i) = \frac{1}{n^2}nVar(X) = \frac{Var(X)}{n}$$

Hence, $\sigma_{\overline{X}} = \frac{\sigma_X}{\sqrt{n}}$

Sampling distribution of mean

- We have defined the mean and variance for the sampling distribution of a mean, but what about the distribution itself?
- Here are boxplots for the earlier example, showing the distributions of the original population and of the population of means from all possible samples of size 3.

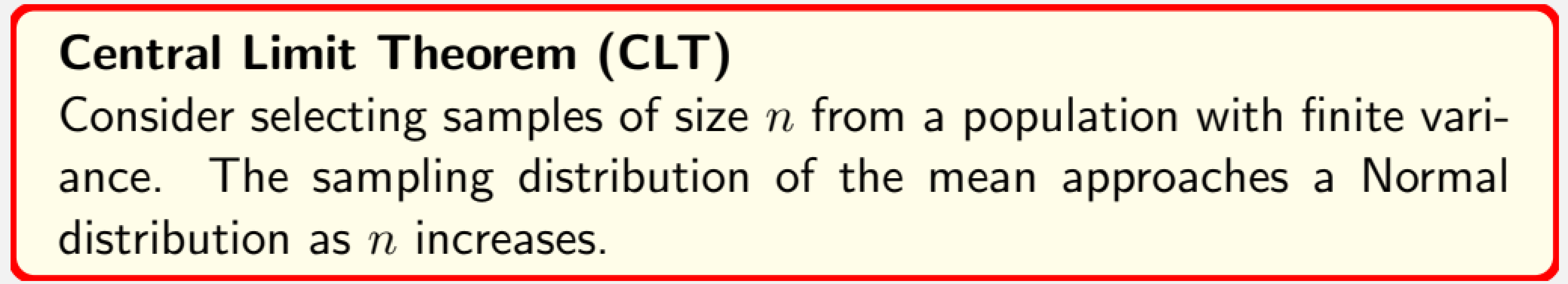


Sampling distribution of mean

- From the results for mean and variance, and now the graphs, it appears that when we compare the sampling distribution of the mean to the population distribution we can say -
  - Location (mean) does not change, and
  - Spread (SD) reduces according to $\frac{1}{\sqrt{n}}$
  - Symmetry increases.
- That is as far as we can go with a little bit of maths and some graphs. It is time for the most important result in statistics ...

Central Limit Theorem

- Using mathematics beyond what we teach in this unit, the following theorem has been proved:



Central Limit Theorem (CLT)

Consider selecting samples of size n from a population with finite variance. The sampling distribution of the mean approaches a Normal distribution as n increases.

Note that -

- CLT works for any population with a finite variance.
- CLT does not specify how large *n* must be for the sampling distribution to be “close enough” to Normal. That is done through simulation studies (such as in Computer Lab 3).
- CLT works *gradually*. It does not suddenly switch on at any particular value of *n*. The effects of averaging are apparent from *n = 2* and the sampling distribution gradually approaches Normality as *n* increases.

Central Limit Theorem

Further notes:

- CLT works more quickly when starting from a roughly symmetric population. Here, $n = 10 - 15$ will probably be enough for the sampling distribution to be very close to Normal.
- For moderately skewed distributions, $n = 25 - 30$ is enough to produce Normal-looking sampling distributions.
- For strongly skewed distributions or distributions with all the weight in the tails (e.g., Bernoulli), $n \geq 100$ may be needed.
- It is common enough for textbooks to say “if $n > 30$ you can rely on
- CLT”. That is roughly true, but you should be aware of the nuances.
- If the population is Normally distributed then sample means will also be Normally distributed, for any sample size.

Sampling distribution of mean

With the Central Limit Theorem added to our earlier results, we can say that the sampling distribution of a mean:

- has $\mu_{\bar{X}} = \mu_X$ and
- Has $\sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{n}}$
- is approximately Normal for sufficiently large n.

$\sigma_{\bar{X}}$ is called the standard error of the sample mean, because we often use the sample mean to estimate the population mean.

Sampling distribution of proportion

CLT also holds for proportions.

- A sample proportion is also a mean - the mean of *n* observations of a Bernoulli(*p*) variable, which takes values of either *0* or *1*.
- That is, a sample proportion is a Binomial variable divided by the sample size *n*.

Therefore, the sampling distribution of a proportion:

- has $\mu_{\hat{p}} = \frac{np}{n} = p$
- has $Var(\hat{p}) = Var\left(\frac{X}{n}\right) = \frac{Var(X)}{n^2} = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}$ and hence,

  $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$ is approximately Normal for sufficiently large *n*.

$\sigma_{\hat{p}}$ is often called the *standard error* of the sample proportion.

Estimation - Point Estimators and Interval Estimators

- A *point estimator* for a parameter is an estimator in the form of a single number without a specified margin of error.
- If a margin of error is provided along with the estimator, then we have an *interval estimator*.
- The interval estimator of a parameter is also known as a *confidence interval* and it takes the form
    - point estimator ± margin of error.
- The sample proportion $p\hat{}$ is a point estimator of the population parameter p.
- Since $\hat{p}$ is a variable, we should provide an interval for a plausible range of values p might take.
- The same applies for $\bar{x}$ as an estimator of μ.

Towards a confidence interval

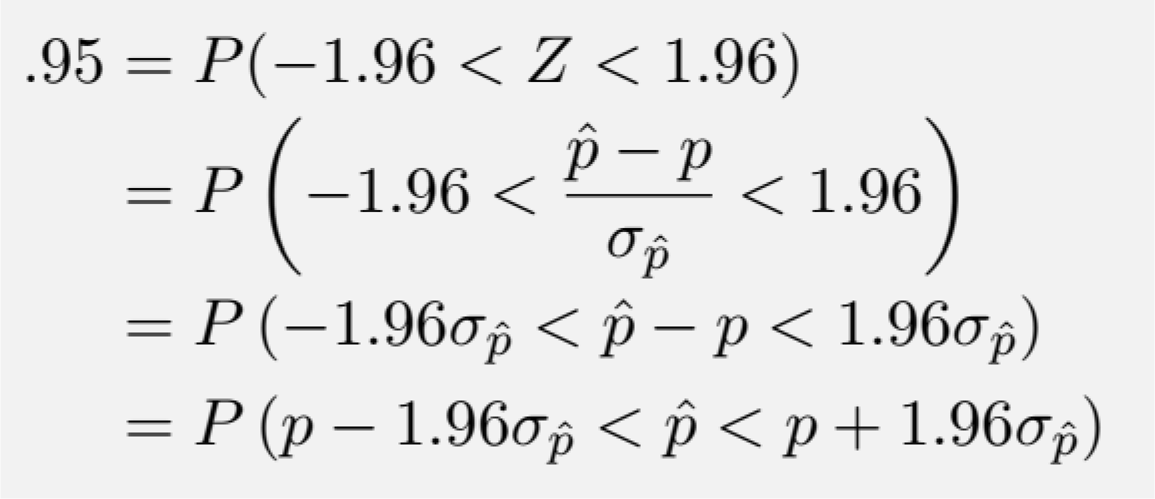
- Suppose we want to determine the probability that a sample proportion
- lies within 0.04 of the population proportion. We could start like this:

$$\begin{aligned} P(p-.04<\hat{p}<p+.04) &= P(-.04<\hat{p}-p<.04) \\ &= P\left(\frac{-.04}{\sigma_{\hat{p}}}<\frac{\hat{p}-p}{\sigma_{\hat{p}}}<\frac{.04}{\sigma_{\hat{p}}}\right) \\ &= P\left(\frac{-.04}{\sigma_{\hat{p}}}<Z<\frac{.04}{\sigma_{\hat{p}}}\right) \end{aligned}$$

- The last line follows since $\hat{p}$~N (p, p(1 − p)/n)if *n* is large.
- The statement is true for the proportion of the sampling distribution that lies within p ± .04. Interpretation as a probability requires that the sample be selected by simple random sampling

Towards a confidence interval

- Alternatively, we could start with a probability and ask, “What are the limits, symmetric around p, within which I can be 95% sure of finding pˆ?”
- We could then start where the previous example finished, replace the quantity $.04/\sigma_{\hat{p}}$ with the 0.025 quantile of a standard Normal distribution (1.96) and work backwards:



$$\begin{aligned}
.95 &= P(-1.96 < Z < 1.96) \\
&= P\left(-1.96 < \frac{\hat{p}-p}{\sigma_{\hat{p}}} < 1.96\right) \\
&= P\left(-1.96\sigma_{\hat{p}} < \hat{p}-p < 1.96\sigma_{\hat{p}}\right) \\
&= P\left(p-1.96\sigma_{\hat{p}} < \hat{p} < p+1.96\sigma_{\hat{p}}\right)
\end{aligned}$$

- Again, probability interpretation depends on SRS.

Towards a confidence interval

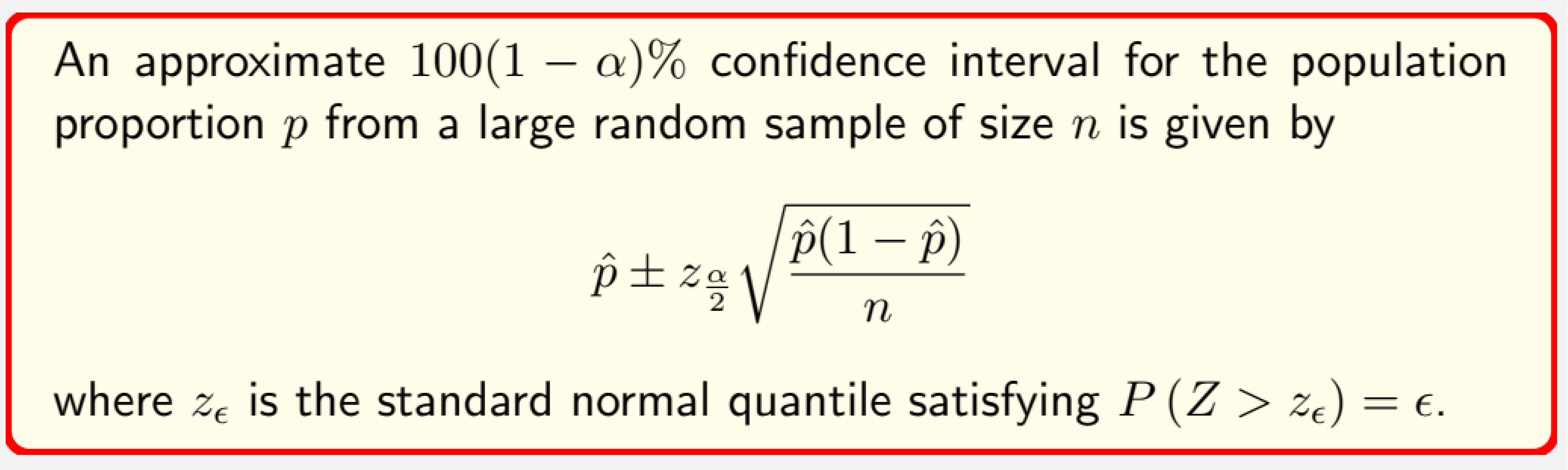
$$.95 = P\left(p - 1.96\sigma_{\hat{p}} < \hat{p} < p + 1.96\sigma_{\hat{p}}\right)$$

- This gives us an interval expression for where we might find $\hat{p}$, assuming we knew p.
- In fact we usually don’t know p, and we want to estimate it using $\hat{p}$ ...
- ... so we turn the expression around and ask, “What are the limits, symmetric around $\hat{p}$, within which I can be 95% sure of finding *p*?”
- The probability expression then becomes

$$.95 = P\left(p - 1.96\sigma_{\hat{p}} < \hat{p} < p + 1.96\sigma_{\hat{p}}\right)$$

- and the limits are therefore $\hat{p} \pm 1.96\sigma_{\hat{p}}$
- This is a 95% confidence interval for a population proportion.

Confidence Interval for p

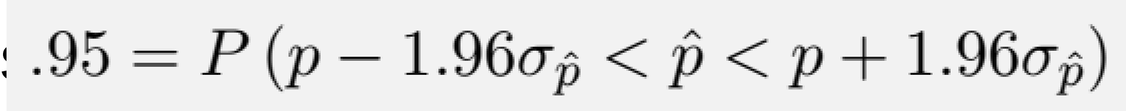


An approximate $100(1-\alpha)\%$ confidence interval for the population proportion p from a large random sample of size n is given by

$$\hat{p} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

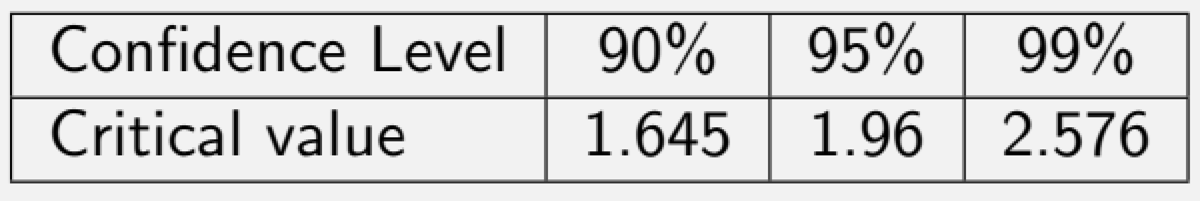
where z_ϵ is the standard normal quantile satisfying $P\left(Z > z_\epsilon\right) = \epsilon$.

- ○ $z_{\alpha/2}$ is known as the normal critical value.
- ○ *1 − α* is known a ce level, which allows for an error probability of α.



$$.95 = P\left(p - 1.96\sigma_{\hat{p}} < \hat{p} < p + 1.96\sigma_{\hat{p}}\right)$$

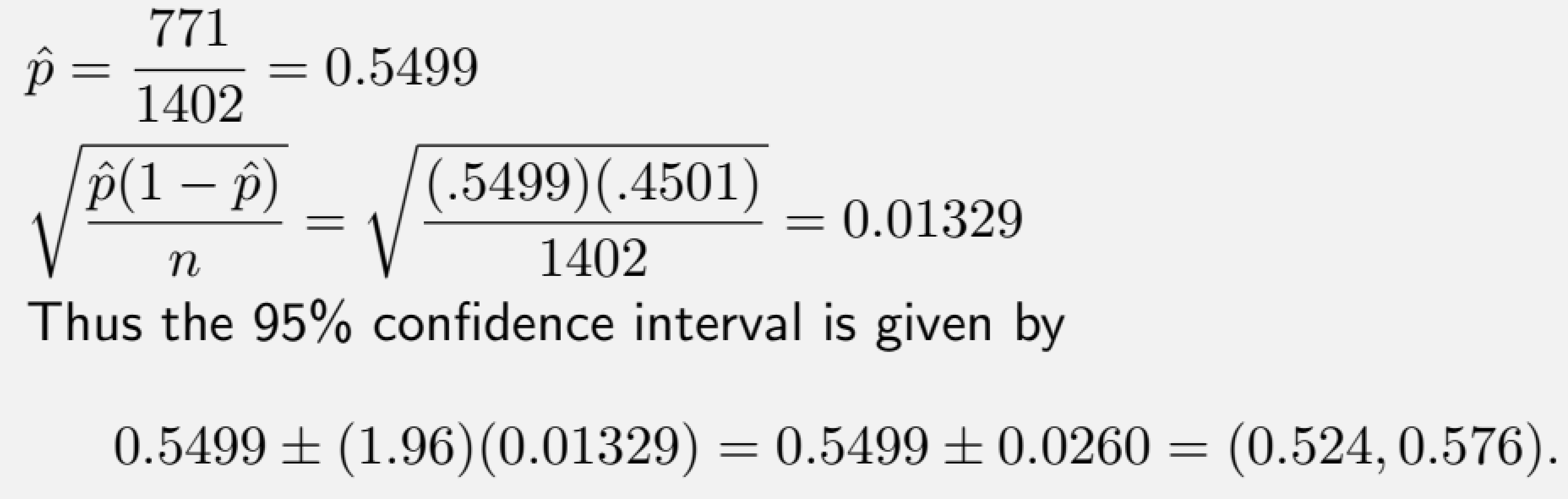
- ○ The most commonly used confidence levels are 0.9, 0.95 and 0.99



Confidence Level	90%	95%	99%
Critical value	1.645	1.96	2.576

A 95% Confidence Interval for p

- **Example**: In a sample of 1402 people, 771 approved of the PM’s performance. Construct a 95% confidence interval for the proportion of voters who approve of the PM’s performance.
- **Solution**:



$$\hat{p} = \frac{771}{1402} = 0.5499$$

$$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{(.5499)(.4501)}{1402}} = 0.01329$$

Thus the 95% confidence interval is given by

$$0.5499 \pm (1.96)(0.01329) = 0.5499 \pm 0.0260 = (0.524, 0.576).$$

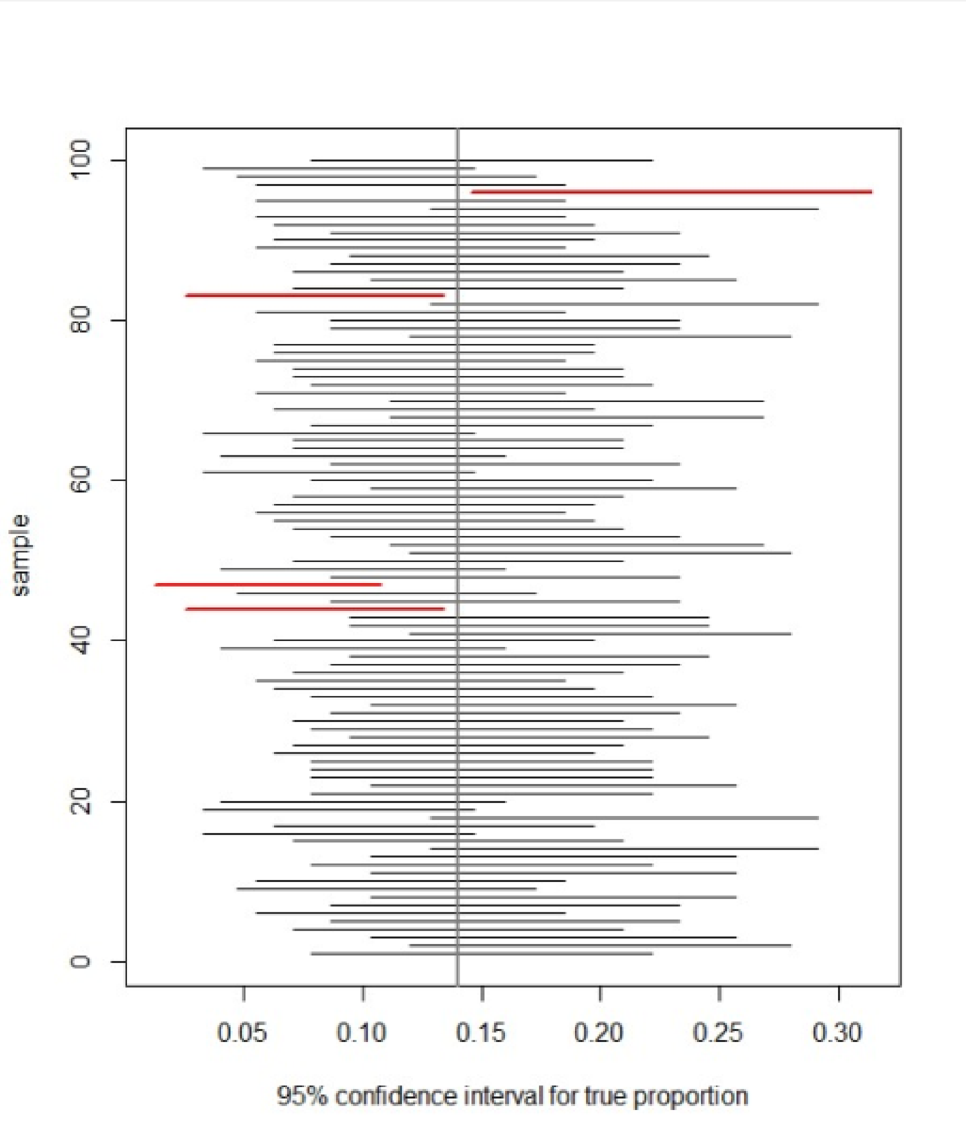
- We conclude with 95% confidence that, if sampling was random, the true proportion p lies between 0.524 and 0.576.

Interpretation of Confidence Intervals

- If we were to do a survey of a fixed size over and over again using many random samples, we'd get many different values of $\hat{p}$, and hence many different confidence intervals.
- The true parameter value lies either inside or outside each such interval.
- The interpretation of a 95% confidence interval is that in these repeated random samples, about 95% of the intervals will contain the true value.
- More generally, the confidence level C is the probability that the interval contains the true parameter value in the sense that if were to repeat the experiment or survey over and over again, C is the fraction of the intervals that would contain the true value of the parameter.

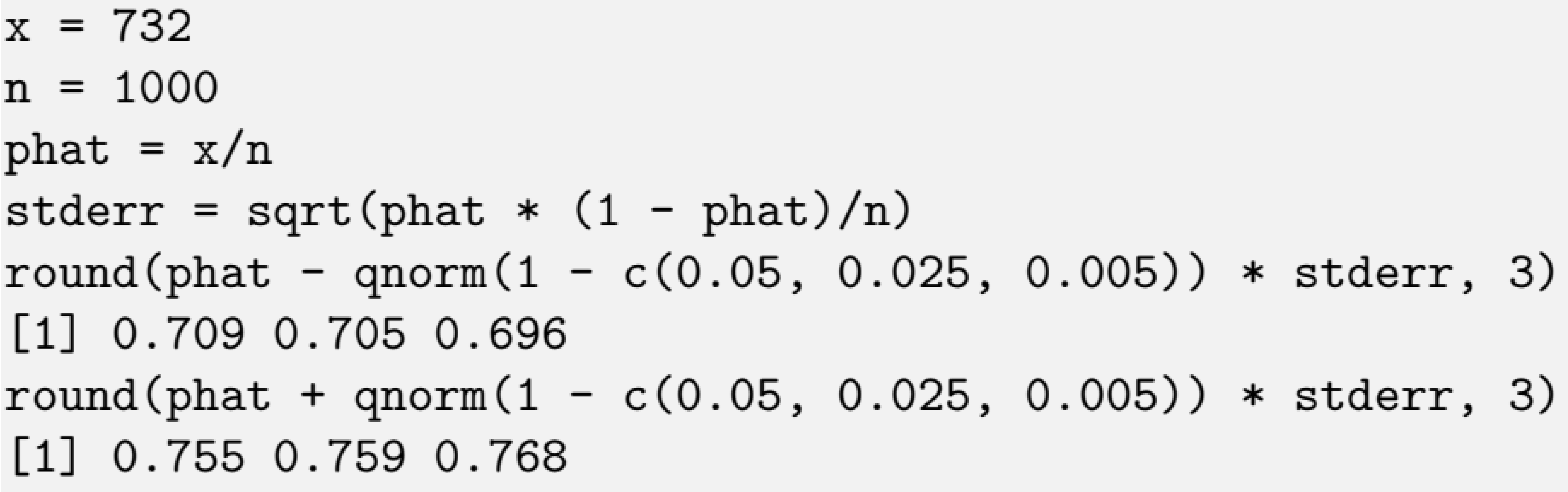
Interpretation of Confidence Intervals

- In the smarties example, suppose the true proportion of yellow smarties is *p = 0.14*
- If we took lots of random samples of the same size, and calculated a 95% confidence interval from each sample, we could expect that about 5% of those intervals would not contain 0.14.



Confidence interval example

- **Example**: An agricultural biochemist testing a new insecticide formulation exposes 1000 mosquitoes to a fixed concentration, and observes that 732 were killed after an exposure period of 2 minutes. Construct 90%, 95%, and 99% confidence intervals for the true proportion of mosquitoes that would be killed.



```
x = 732
n = 1000
phat = x/n
stderr = sqrt(phat * (1 - phat)/n)
round(phat - qnorm(1 - c(0.05, 0.025, 0.005)) * stderr, 3)
[1] 0.709 0.705 0.696
round(phat + qnorm(1 - c(0.05, 0.025, 0.005)) * stderr, 3)
[1] 0.755 0.759 0.768
```

- The outputs are respectively the lower limits and upper limits of the intervals. Thus the 90%, 95%, and 99% confidence intervals are, respectively, (.709, .755), (.705, .759), and (.696, .768).

How Confidence Intervals Behave

Some important properties are shared by all confidence intervals:

- There is always a trade-off between confidence level and margin of error. The user chooses the confidence level, carries out the survey or experiment and the margin of error follows.
- Ideally, we want high confidence and a small margin of error. However,
    - For given data, a high confidence level will produce a larger margin of error (wider interval).
    - A lower confidence level will allow a smaller margin of error.
- The margin of error for p is $z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ which gets smaller when:
    - $z_{\alpha/2}$ gets smaller (reflecting a lower confidence level).
    - n gets larger.
- Since *n* is under the square root sign, we would need four times as many observations to cut the margin of error in half.

Choosing the Sample Size

- We can find the required sample size for a desired margin of error $e = z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ to get:

$$n = \left(\frac{z_{\frac{\alpha}{2}}}{e}\right)^2 p^*(1-p^*)$$

- where p^* is a guessed value for p. The following guidelines can be used for the choice of p^* in the given order:
  1. If a previously estimated value of p is available, use it.
  2. If a range of possible values for p is available, use the value that is closest to 0.5 within the range.
  3. If no information whatsoever is available, use $p^* = 0.5$

Sample size examples

- We always round up to the nearest integer. For instance, 35.3 will be rounded to 36. This is done to make sure that the margin of error is no more than e.

- **Example**: An agricultural biochemist would like to estimate with 95% confidence the proportion of insects that would be affected by a fixed concentration of a particular insecticide with a margin of error no greater than 3%. How large a sample does she need in her test?
- Here no information about p is available, so

$$n = \left(\frac{z_{\frac{\alpha}{2}}}{e}\right)^2 p^*(1-p^*) = \left(\frac{1.96}{.03}\right)^2 \left(\frac{1}{2}\right)\left(\frac{1}{2}\right) = 1067.1.$$

- We round this up to n = 1068.

Sample size examples

- **Example**: A health agency is trying to estimate the proportion of
- smokers among high school students in a city. From past data, the
- agency believes the proportion is about 0.2. Find the required sample
- size if the margin of error for a 90% confidence interval is to be no
- more than 0.04.
- Here we can take p^* to be 0.2 to get

$$n = \left(\frac{1.645}{.04}\right)^2 (.2)(.8) = 270.6$$

rounded to 271.

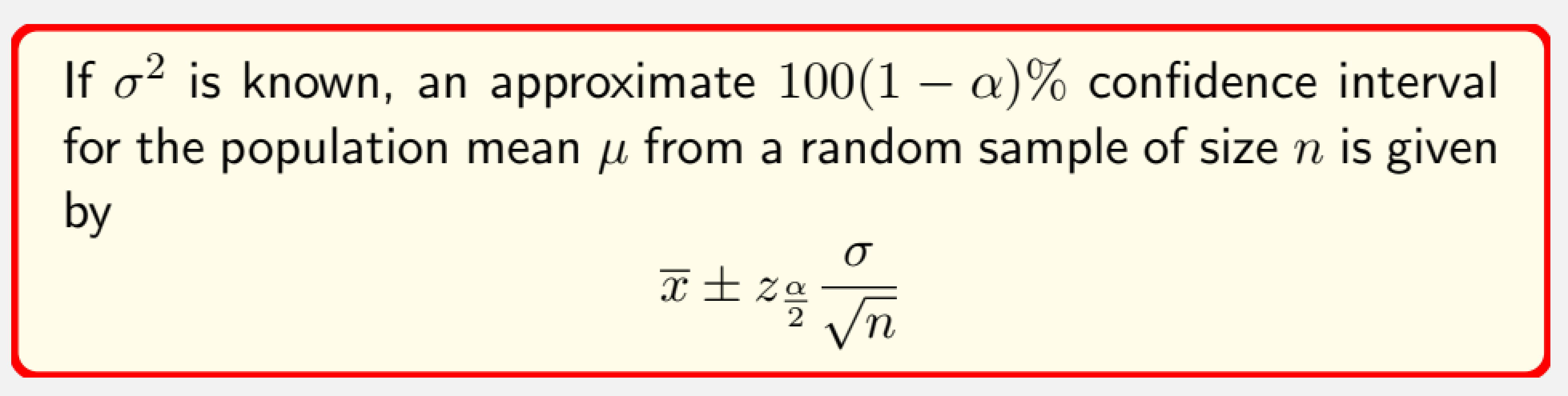
- Note that without the information about possible values of p, we would have used

$$n = \left(\frac{1.645}{.04}\right)^2 (.5)(.5) = 422.8,$$

- which meant we would have been forced to take 423 observations, quite a bit more than 271.

Examples

- Suppose we want to construct a confidence interval for the population mean μ. For now, assume that the population variance σ^2 is known.



If σ^2 is known, an approximate $100(1-\alpha)\%$ confidence interval for the population mean μ from a random sample of size n is given by

$$\overline{x} \pm z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

- If the population distribution is normal, the confidence interval is exact, whether or not *n* is large.
- **Interpretation**: Comments made earlier about general properties of
- confidence intervals and their interpretation also apply here.

Confidence Intervals for μ

- **Example**: A random sample of size 50 was taken from the adult male population of a country and the weights (in kg) of the selected men are obtained. The population standard deviation is known to be 10 kg. If the sample mean was found to be 75 kg, construct 90% and 99% confidence intervals for the mean weight of all adult males.
- The *100(1 − α)%* confidence interval is given by $75 \pm z_{\alpha/2}\frac{10}{\sqrt{50}}$, so the
- 90% confidence interval is

$$75 \pm 1.645\frac{10}{\sqrt{50}} = 75 \pm 2.33 = (72.67, 77.33)\ \text{kg}$$

and the 99% confidence interval is

$$75 \pm 2.576\frac{10}{\sqrt{50}} = 75 \pm 3.64 = (71.36, 78.64)\ \text{kg}$$

When σ^2 is Unknown

- In fact, the population variance is rarely known. When σ^2 is unknown, we are obliged to use its estimator s^2, the sample variance, and this changes the relevant sampling distribution.
- The sample variance is given by the formula

$$s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\right)$$

- When σ^2 is known the standardised sample mean $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim$ N(0, 1).
- This is exactly true if $X \sim N(\mu, \sigma^2)$ and approximately true otherwise if *n* is large.
- If we replace *σ* with *s*, the normal distribution statement is no longer true (though it would still be approximately true if *n* were large enough).

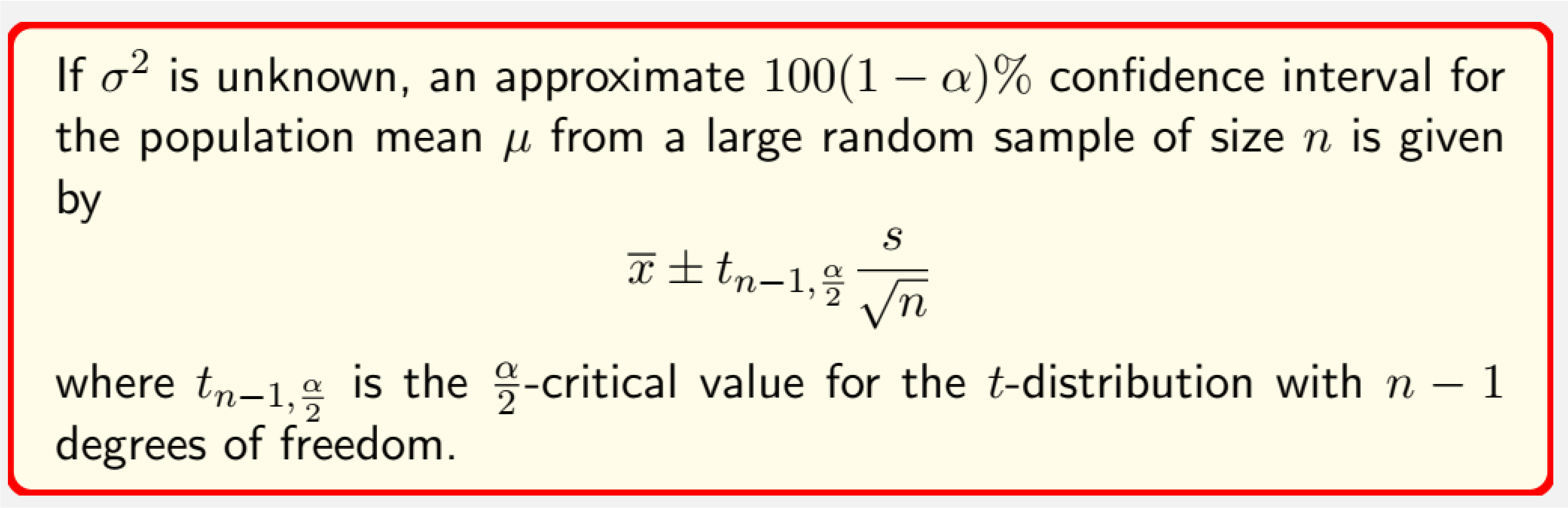
Student's *t*-distribution

- In these circumstances, we need to use a distribution known as the Student's t-distribution.
- If X is Normally distributed, then

$$T = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim t_{n-1}$$

- which is a t distribution with *n* − *1* degrees of freedom.
- The "degrees of freedom" directly reflects the *n* − *1* denominator in the calculation of the sample standard deviation *s*.
- A *t* distribution is like a Standard Normal distribution, but with thicker tails. This reflects the greater variability created by having two estimates ($\bar{x}$ and *s*) in the standardised score.
- As the degrees of freedom increase, the t-distribution approaches the Standard Normal.

A *t*-Confidence Interval



If σ^2 is unknown, an approximate $100(1-\alpha)\%$ confidence interval for the population mean μ from a large random sample of size n is given by

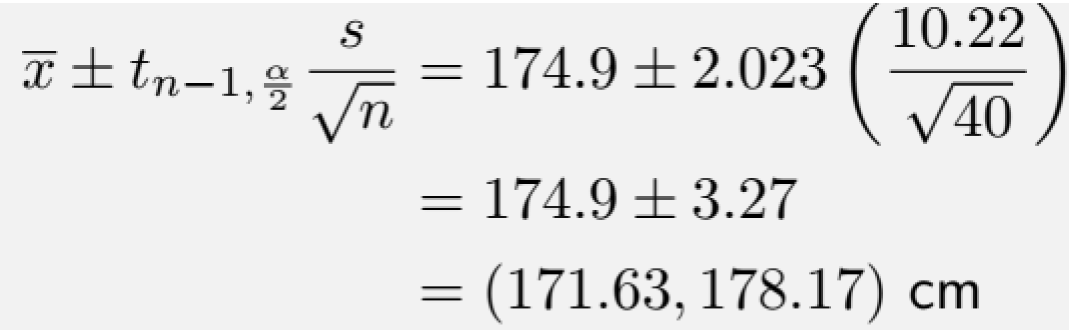
$$\overline{x} \pm t_{n-1,\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$$

where $t_{n-1,\frac{\alpha}{2}}$ is the $\frac{\alpha}{2}$-critical value for the t-distribution with $n-1$ degrees of freedom.

- If the population distribution is normal, the confidence interval is exact, whether or not n is large.
- The method is acceptable for $n \geq 20$ if there are no outliers and the population distribution is only mildly skewed, and for $n \geq 40$ if there are no outliers and the skew is stronger.

Confidence interval for μ - examples

- **Example**: The heights of 40 randomly selected women are measured, and
- give $\bar{x} = 174.9$ and $s^2 = 104.5$. Calculate the 95% confidence interval for
- the population mean μ.
- **Solution**: As $s^2 = 104.5$, we have *s* = √*104.5* = *10.22*. From R or the
- tables, we get $t_{39,0.025} = 2.023$. Thus, the confidence interval is given by



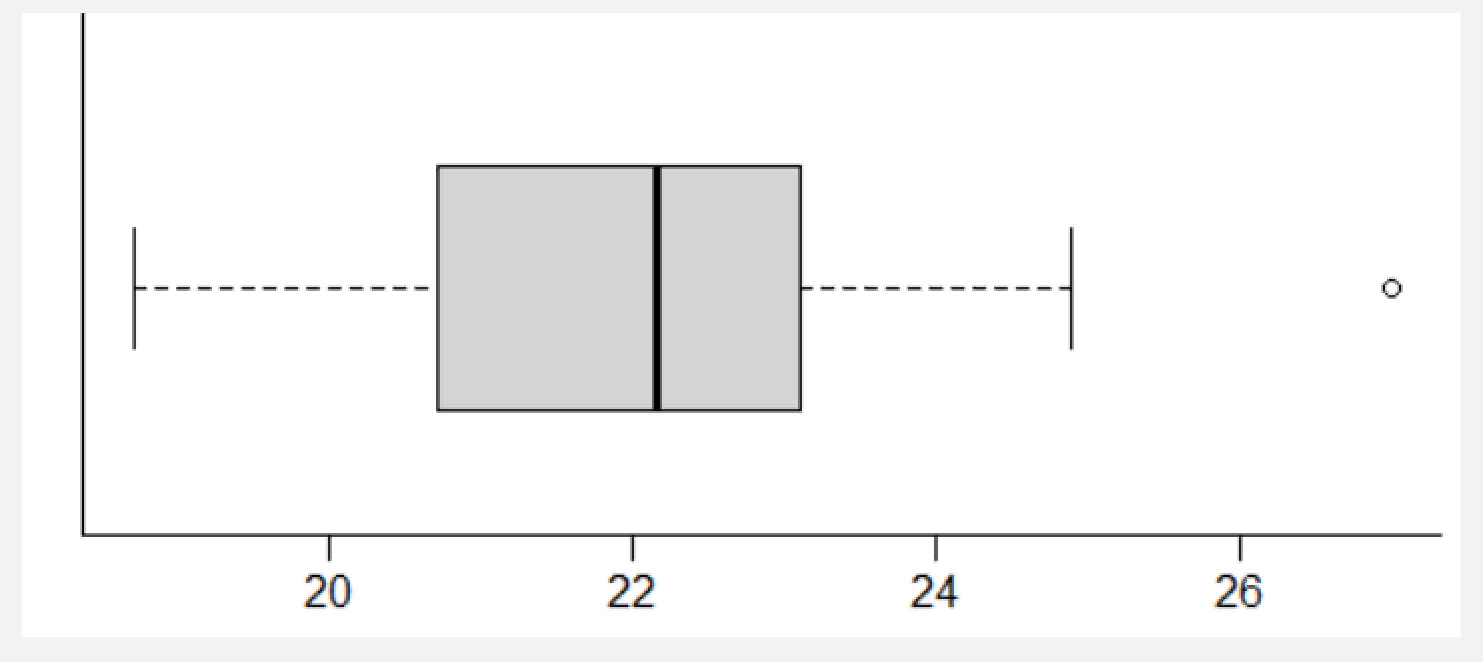
$$\begin{aligned}\overline{x} \pm t_{n-1,\frac{\alpha}{2}} \frac{s}{\sqrt{n}} &= 174.9 \pm 2.023 \left(\frac{10.22}{\sqrt{40}} \right) \\ &= 174.9 \pm 3.27 \\ &= (171.63, 178.17) \text{ cm}\end{aligned}$$

- The R command for obtaining the critical value is `qt(.975,39)`.
- If we have to use the table to find $t_{n-1,\,\frac{\alpha}{2}}$ and it does not list the degrees of freedom we need, we can approximate using the nearest value.

Confidence interval for μ - examples

- Example: The following represents cholesterol levels in a random sample of
- students. Plot the data and calculate a 90% confidence interval for the true
- mean cholesterol level.

20.8 18.7 19.9 20.6 21.9 23.4 22.8
24.9 22.2 20.3 24.9 22.3 27.0 20.4
22.2 24.0 21.1 22.1 22.0 22.7



Confidence interval for µ - examples

- The data are roughly symmetric except for a high outlier indicated on the boxplot. In view of the outlier we should treat our results with some caution.
- Descriptive statistics are $\bar{x} = 22.21$ and $s = 1.962$.
- As *n* = *20*, *df* = *19*, and $t_{19,0.05} = 1.729$.
- Thus the 90% confidence interval for μ is given by

$$\begin{aligned}\overline{x} \pm t_{n-1,\frac{\alpha}{2}} \frac{s}{\sqrt{n}} &= 22.21 \pm 1.729 \left(\frac{1.962}{\sqrt{20}}\right) \\ &= 22.21 \pm 0.759 \\ &= (21.45, 22.97)\end{aligned}$$