8. Sampling distributions and confidence intervals

1

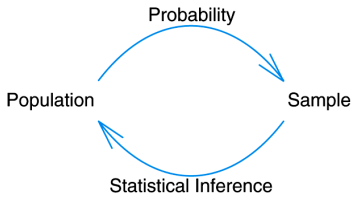
Learning Outcomes

- Explain sampling variation, sampling distribution, standard error
- Calculate the standard error of the sample mean
- State the Central Limit Theorem (applied to sampling distribution of the sample mean)
- Describe the sampling distribution of the sample mean in applications using the CLT
- Identify the point estimator of the parameter in applications
- Describe briefly the use of a confidence interval in inferential statistics
- Calculate and interpret 95% confidence interval for the population mean
- Use R to calculate the standard error and calculate a 95% confidence interval for the population mean
- Use the t distribution to calculate the standard error and confidence intervals for the population mean using a small sample
- Confidence intervals for the mean and other statistics via simulation, using R

1 - 2

2

Fundamental relationship between probability and inferential statistics



3

3

Probability and Statistics

- In probability theory we consider some **known process** which has some randomness or uncertainty. We model the outcomes by random variables, and we figure out the probabilities of what will happen. There is one correct answer to any probability question.
- In statistical inference we observe something that has happened, and try to **figure out what underlying process** would explain those observations.

4

4

An example …

- Consider an (opaque) jar of red and green jelly beans.
- A probabilist starts by **knowing the proportion** of each and asks: What is the probability of drawing a red jelly bean from the jar?
- A statistician **infers the proportion** of red jelly beans by sampling from the jar, and using the sample proportion to estimate the jar proportion.

5

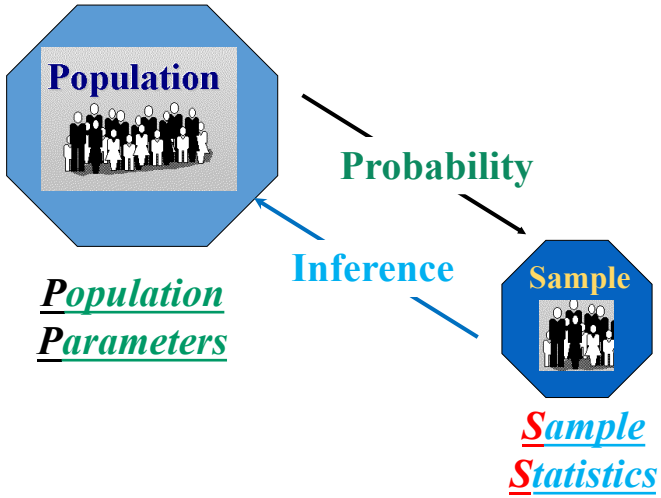
5

Probability and Statistics

- The basic aim behind all statistical methods is to make inferences about a population by studying a relatively small sample chosen from it.
- Probability is the engine that drives all statistical modelling, data analysis and inference.

6

6



7

7

Foundations for Inference

- Recall that inference is concerned (primarily) with estimating population parameters using sample statistics.
- A classic inferential question is, "How sure are we that the sample mean, x, is near the true population mean, μ?"
- Estimates (i.e. statistics) generally vary from one sample to another, and an understanding of **sampling variation** is key when estimating the precision of a sample statistic as an estimate of the corresponding parameter.

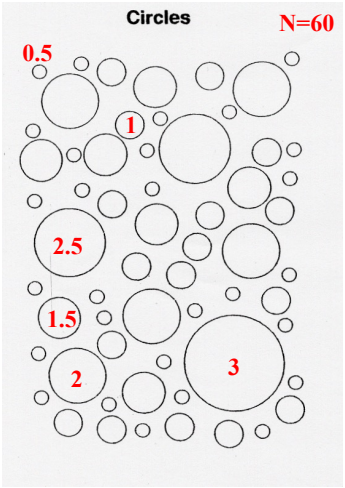
8

8

Sampling Distributions

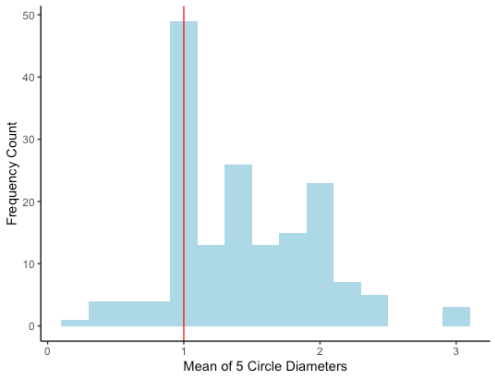
- The probability distribution of a statistic is called a sampling distribution.
- Sampling distributions arise because samples vary.
- Each random sample will have a different value of the statistic.

9

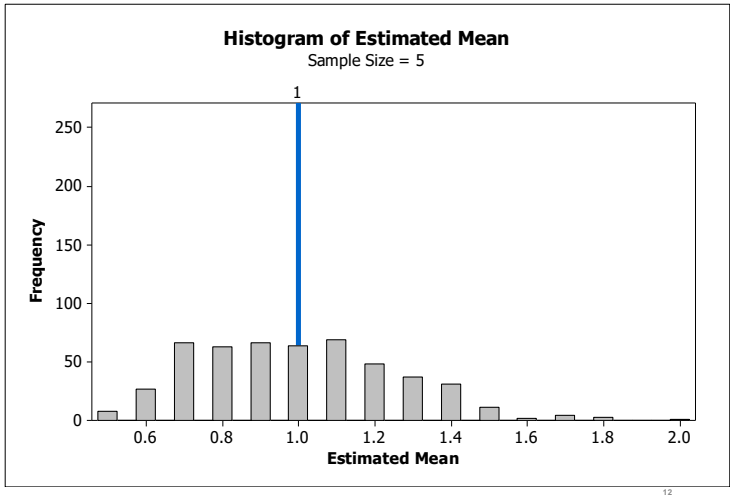


10

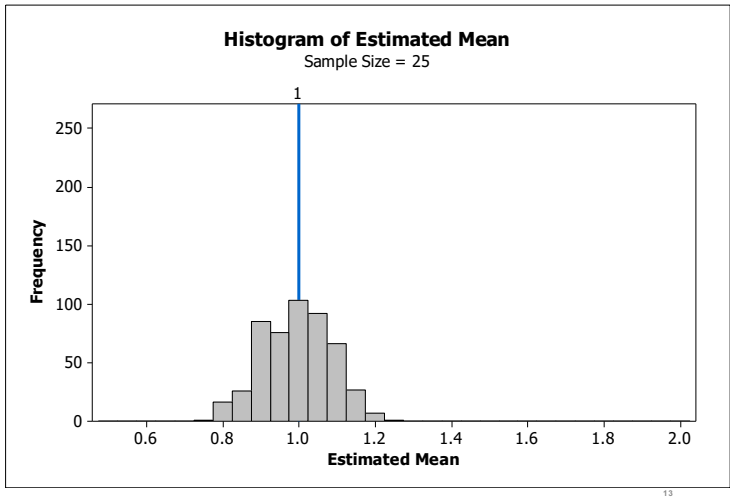
Judgement Sample



11

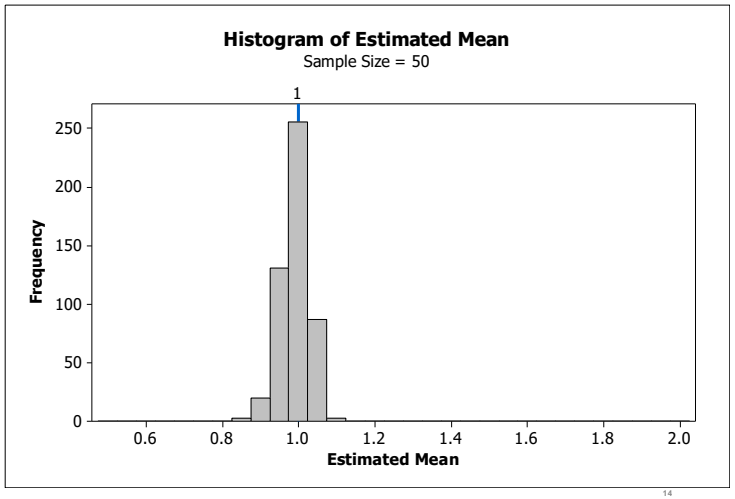


12



13

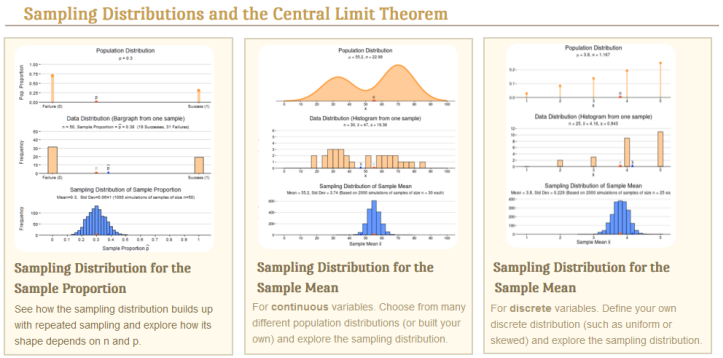
13



14

14

http://www.artofstat.com/webapps.html



15

15

The Central Limit Theorem

- The sampling distribution of *any* mean becomes more nearly Normal as the sample size grows
  - observations need to be independent.
  - the shape of the population distribution doesn't matter.

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

16

16

The Central Limit Theorem

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

The CLT depends crucially on the assumption of independence.

You can't check this with your data. You have to think about how the data were gathered – can you assume the observations are independent?

1 - 17

17

The Central Limit Theorem

- *Sample means follow a Normal distribution centred on the population mean with a standard deviation equal to population standard deviation divided by the square root of the sample size.*

$$\bar{\boldsymbol{X}} \sim \boldsymbol{N}\left(\boldsymbol{\mu}, \frac{\boldsymbol{\sigma^2}}{\boldsymbol{n}}\right)$$

- What happens when you take a single sample ?

18

18

The Standard Error

- The standard error is a measured of the variability in the sampling distribution (i.e. how do sample statistics vary about the unknown population parameter they are trying to estimate)
- It describes the typical 'error' or 'uncertainty' associated with the estimate.

$$\bar{\boldsymbol{X}} \sim \boldsymbol{N}\left(\boldsymbol{\mu}, \frac{\boldsymbol{\sigma^2}}{\boldsymbol{n}}\right) \qquad SE = \frac{\sigma}{\sqrt{n}}$$

1 - 19

19

Interval Estimation for μ

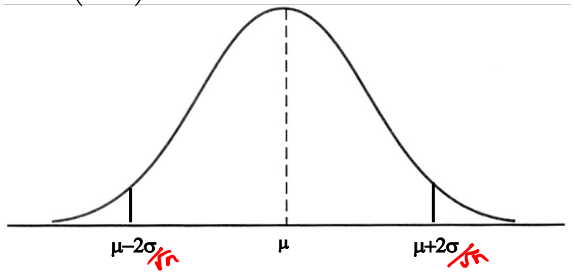
Use the CLT to provide a range of values that will capture 95% of sample means.

20

20

95% of sample means

$$X \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

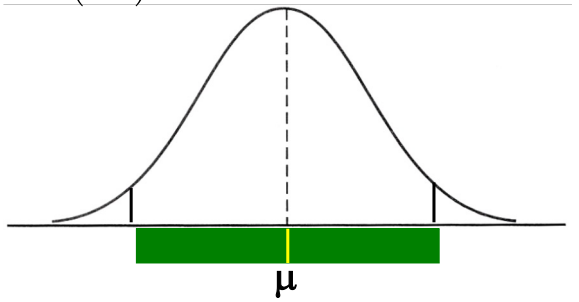


21

21

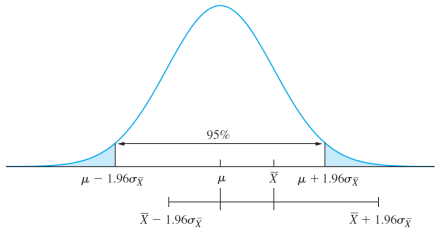
95% of sample means

$$X \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$



22

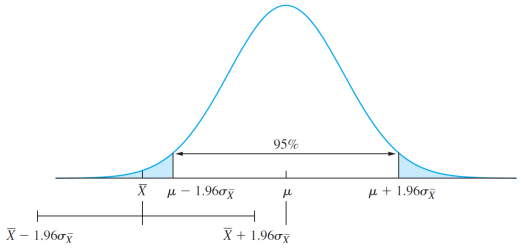
22



The sample mean X has a normal distribution with mean μ and standard deviation $\sigma_X = \sigma/\sqrt{n}$.
Let's consider a particular sample with mean x.
Now suppose x lies in the middle 95% of the distribution of X — the 95% confidence interval $x \pm 1.96\sigma_X$ succeeds in covering the population mean μ.

23

23



The sample mean X has a normal distribution with mean μ and standard deviation $\sigma_X = \sigma/\sqrt{n}$.
Let's consider a particular sample with mean x.
Now suppose x lies in the outer 5% of the distribution of X — the 95% confidence interval $x \pm 1.96\sigma_X$ does not include the population mean μ.

24

24

95% Confidence Interval for μ

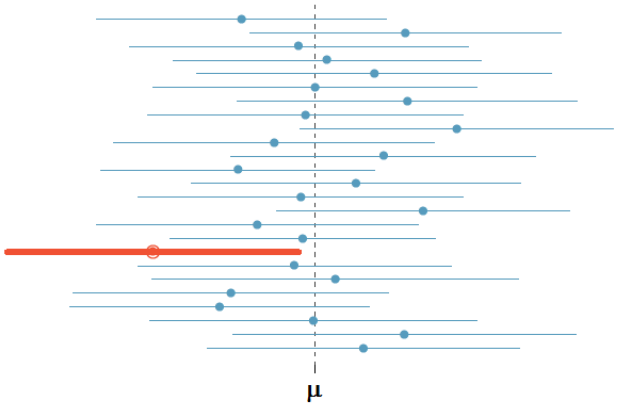
In repeated sampling, 95% of intervals calculated in this manner

$$\bar{x} \pm 1.96 * \frac{\sigma}{\sqrt{n}}$$

will contain the true mean μ.

25

25



26

26

Confidence intervals

- The population mean μ is **fixed**
- The intervals from different samples are **random**
- From our single sample, we only observe one of the intervals
- Our interval may or may not contain the true mean
- If we had taken many samples and calculated the 95% CI for each, 95% of them would include the true mean
- We say we are "95% confident" that the interval contains the true mean.

1 - 27

27

Confidence Intervals

- A point estimate (i.e. a statistic) is a single plausible value for a parameter.
- A point estimate is rarely perfect; usually there is some error in the estimate.
- Instead of supplying just a point estimate of a parameter, a next logical step would be to provide a plausible range of values for the parameter.
- To do this an estimate of the precision of the sample statistic (i.e. the estimate) is needed.

28

28

$n = 150, \bar{x} = 69.5, \sigma = 6.2$

Is $\mu > 75$?

29

$$69.5 \pm 1.96 \frac{6.2}{\sqrt{150}}$$

```
> 69.5-1.96*6.2/sqrt(150)
[1] 68.50779
> 69.5+1.96*6.2/sqrt(150)
[1] 70.49221
```

A 95% CI for the population mean is [68.51, 70.49]

Interpret this !

Is $\mu > 75$?

30

95% confident that the population mean is between 68.48 and 70.51 based on the data provided.

No evidence to support the claim that the population mean (μ) greater than 75.

31

Application: mean weekly rent in ST2001

```
survey.data %>%
  select(rent) %>%
   filter(rent>0 & rent < 5000) %>%
  summarise(sample.size = n(),
            mean = mean(rent),
            sd = sd(rent))
```

```
##   sample.size     mean       sd
## 1         108 617.8056 214.7341
```

What is the population mean rent ?
What is a student likely to pay ?
What will they actually pay ?

32

Population Mean Rent in ST2001 ?

```
survey.data %>%
  select(rent) %>%
  filter(rent>0 & rent < 5000) %>%
  t.test()
```

```
##
##  One Sample t-test
##
## data:  .
## t = 29.899, df = 107, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  576.8440 658.7671
## sample estimates:
## mean of x
##  617.8056
```

33

Using s for σ ?

- Knowing s must mean that you knew μ ....

$$\bar{x} \pm 1.96 * \frac{\sigma}{\sqrt{n}}$$

- The sample standard deviation s is used to estimate σ.
- What are the consequences ?

34

What if σ is unknown and n is small ?

35

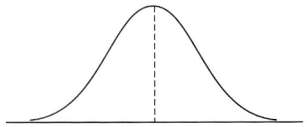
$$n < 30$$

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

1- confidence level

Degrees of free

Population normal



36

t_ν distribution

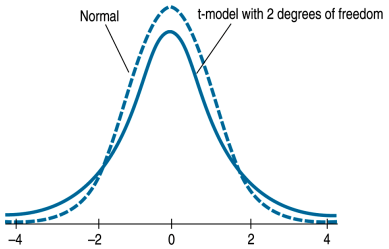
Mean = 0

Variance = $\frac{\nu}{\nu-2}$ for $\nu > 2$

1 - 37

37

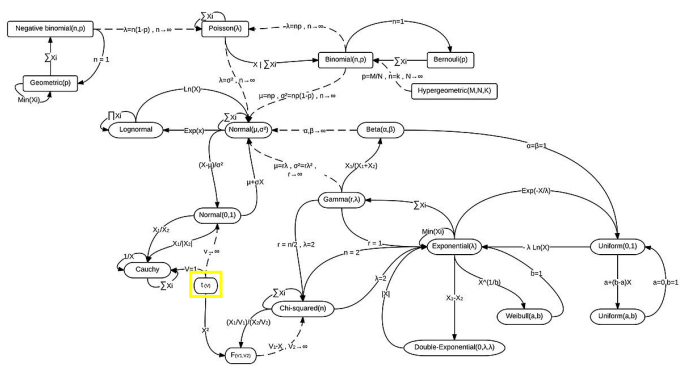
T- distribution



- As the degrees of freedom increase, the *t*-models look more and more like the Normal.
- In fact, the *t*-distribution with infinite degrees of freedom is the Normal distribution.

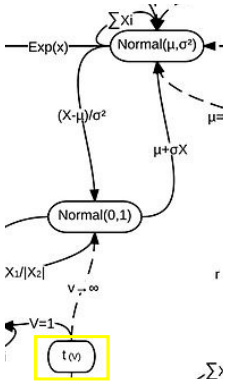
38

38



39

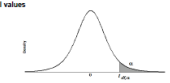
39



40

40

Table: t distribution critical values



T- tables

	Upper tail probability											
df	0.25	0.20	0.15	0.10	0.05	0.025	0.02	0.01	0.005	0.0025	0.001	0.0005
1	1.000	1.376	1.963	3.078	6.314	12.706	15.895	31.821	63.657	127.321	318.309	636.619
2	0.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.089	22.327	31.599
3	0.765	0.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.215	12.924
4	0.741	0.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025	4.437
12	0.695	0.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930	4.318
13	0.694	0.870	1.079	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852	4.221
14	0.692	0.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140
15	0.691	0.866	1.074	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.073
16	0.690	0.865	1.071	1.337	1.746	2.120	2.235	2.583	2.921	3.252	3.686	4.015
17	0.689	0.863	1.069	1.333	1.740	2.110	2.224	2.567	2.898	3.222	3.646	3.965
18	0.688	0.862	1.067	1.330	1.734	2.101	2.214	2.552	2.878	3.197	3.610	3.922
19	0.688	0.861	1.066	1.328	1.729	2.093	2.205	2.539	2.861	3.174	3.579	3.883
20	0.687	0.860	1.064	1.325	1.725	2.086	2.197	2.528	2.845	3.153	3.552	3.850
21	0.686	0.859	1.063	1.323	1.721	2.080	2.189	2.518	2.831	3.135	3.527	3.819
22	0.686	0.858	1.061	1.321	1.717	2.074	2.183	2.508	2.819	3.119	3.505	3.792
23	0.685	0.858	1.060	1.319	1.714	2.069	2.177	2.500	2.807	3.104	3.485	3.768
24	0.685	0.857	1.059	1.318	1.711	2.064	2.172	2.492	2.797	3.091	3.467	3.745
25	0.684	0.856	1.058	1.316	1.708	2.060	2.167	2.485	2.787	3.078	3.450	3.725
26	0.684	0.856	1.058	1.315	1.706	2.056	2.162	2.479	2.779	3.067	3.435	3.707
27	0.684	0.855	1.057	1.314	1.703	2.052	2.158	2.473	2.771	3.057	3.421	3.690
28	0.683	0.855	1.056	1.313	1.701	2.048	2.154	2.467	2.763	3.047	3.408	3.674
29	0.683	0.854	1.055	1.311	1.699	2.045	2.150	2.462	2.756	3.038	3.396	3.659
30	0.683	0.854	1.055	1.310	1.697	2.042	2.147	2.457	2.750	3.030	3.385	3.646
40	0.681	0.851	1.050	1.303	1.684	2.021	2.123	2.423	2.704	2.971	3.307	3.551
50	0.679	0.849	1.047	1.299	1.676	2.009	2.109	2.403	2.678	2.937	3.261	3.496
60	0.679	0.848	1.045	1.296	1.671	2.000	2.099	2.390	2.660	2.915	3.232	3.460
80	0.678	0.846	1.043	1.292	1.664	1.990	2.088	2.374	2.639	2.887	3.195	3.416
100	0.677	0.845	1.042	1.290	1.660	1.984	2.081	2.364	2.626	2.871	3.174	3.390
1000	0.675	0.842	1.037	1.282	1.646	1.962	2.056	2.330	2.581	2.813	3.098	3.300
Z^1	0.674	0.842	1.036	1.282	1.645	1.960	2.054	2.326	2.576	2.807	3.090	3.291

41

41

Table: t distribution critical values

T- tables

	Upper tail probability											
df	0.25	0.20	0.15	0.10	0.05	0.025	0.02	0.01	0.005	0.0025	0.001	0.0005
1	1.000	1.376	1.963	3.078	6.314	12.706	15.895	31.821	63.657	127.321	318.309	636.619
2	0.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.089	22.327	31.599
3	0.765	0.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.215	12.924
4	0.741	0.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587

42

42

One-sample t-interval for a population mean

- When the conditions are met, we are ready to find the confidence interval for the population mean, μ.
- The confidence interval is

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

- The critical value $t_{(1-\frac{\alpha}{2},n-1)}$ depends on the particular confidence level, *1- α*, that you specify and on the number of degrees of freedom, $n - 1$, which we get from the sample size.
- **Let R do the work ….**

43

43

Example: Celtic study

44

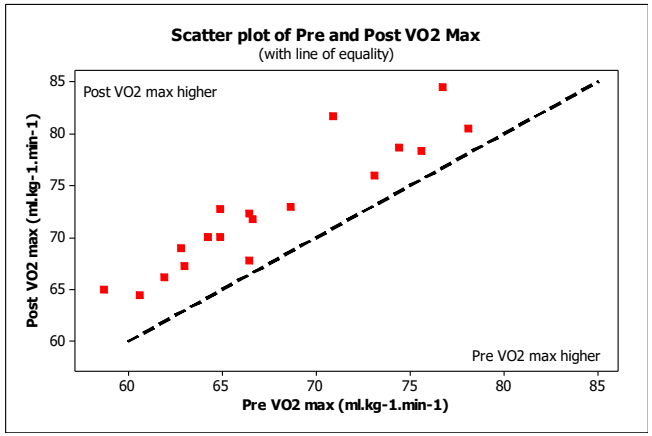
44

Celtic Study

- A sample of 18 full-time youth soccer players from a Youth Academy performed high intensity aerobic interval training over a 10-week in-season period ***in addition*** to usual regime of soccer training and matches.
- Did this extra training improve fitness (VO2 max) ?
- Paired design: each player measured before and after (i.e. start and after 10 weeks)

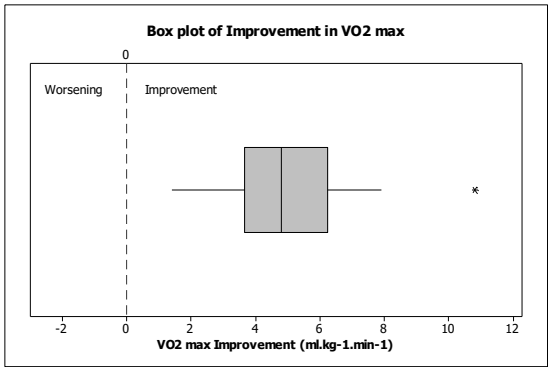
45

45



46

46



Variable	N	Mean	StDev
VO2 Improvement	18	5.11111	2.25829

47

47

Estimate the population mean improvement

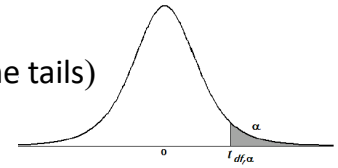
- 90% CI for μ
- 95% CI for μ $\quad \bar{x} \pm t_{(1-\frac{\alpha}{2},n-1)} \frac{s}{\sqrt{n}}$
- 99% CI for μ

48

48

Estimate the population mean improvement

- 90% CI for μ
  ($\alpha = 0.10$ split over the tails)
- 95% CI for μ
  ($\alpha = 0.05$ split over the tails)



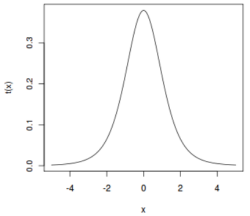
- 99% CI for μ ($\alpha = 0.10$)
  $\alpha = 0.01$ split over the tails

49

49

Using R to calculate the quantile needed corresponding to a particular tail area

The qt(p=? , df= ?, lower.tail=TRUE) function calculates the t-value corresponding to a given lower-tailed area.



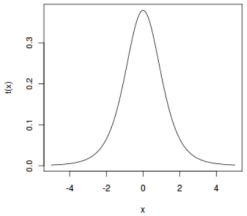
- Find the percentile of the Student t distribution needed for a 95% CI from a sample of size 18.

50

50

- Find the percentile of the Student t distribution needed for a 95% CI from a sample of size 18.

- For a 95% CI need the percentiles corresponding to tail areas such that 95% of the distribution is between these percentiles (i.e. 5% of the area split across the two tails).



- To calculate the 2.5th and 97.5th percentiles of the Student t distribution with 17 degrees of freedom:

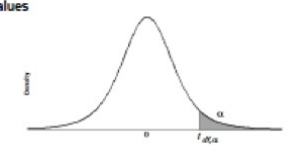
```
> qt(0.975, df=17)
[1] 2.109816
```

51

51

Check the tables …

Table: t distribution critical values



df	Upper tail probability											
	0.25	0.20	0.15	0.10	0.05	0.025	0.02	0.01	0.005	0.0025	0.001	0.0005
14	0.692	0.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140
15	0.691	0.866	1.074	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.073
16	0.690	0.865	1.071	1.337	1.746	2.120	2.235	2.583	2.921	3.252	3.686	4.015
17	0.689	0.863	1.069	1.333	1.740	2.110	2.224	2.567	2.898	3.222	3.646	3.965
18	0.688	0.862	1.067	1.330	1.734	2.101	2.214	2.552	2.878	3.197	3.610	3.922

1 - 52

52

Estimate the population mean improvement

•95% CI for μ $\bar{x} \pm t_{(1-\frac{\alpha}{2}, n-1)} \frac{s}{\sqrt{n}}$

Variable	N	Mean	StDev
VO2 Improvement	18	5.11111	2.25829

```
> qt(0.975, df=17)
[1] 2.109816
```

53

Lower 95% CI using summary statistics

```{r}
5.11 - qt(0.975, df=17)*(2.25829/sqrt(18))
```

```
[1] 3.986979
```

Upper 95% CI using summary statistics

```{r}
5.11 + qt(0.975, df=17)*(2.25829/sqrt(18))
```

```
[1] 6.233021
```

54

Using the t.test function

```{r}
train.df %>% select(Improvement) %>% t.test()
```

```
        One Sample t-test

data:  .
t = 9.6022, df = 17, p-value = 2.798e-08
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 3.988090 6.234132
sample estimates:
mean of x
 5.111111
```

55

Conclusion ?

- On average ?
- What does 95% Confidence mean ?
- Terms and conditions ?
- Random sample ?
- Small n, normality ??

56

If Normality is questionable

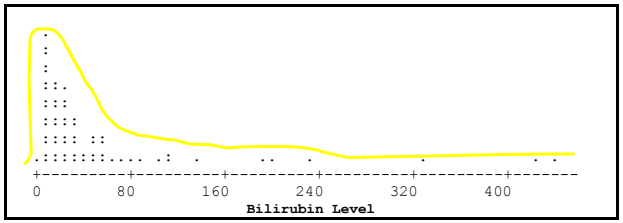
a) Try to transform the data to approximate Normality
   - e.g. logarithms or square root

b) Non-Parametric technique
   - Bootstrap
   - CI for the population MEDIAN

57

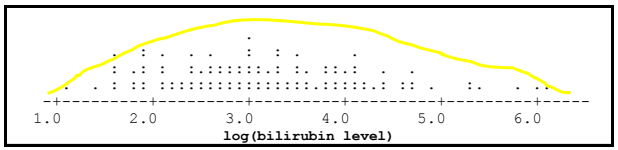
Transforming to Normality

- ***Example:*** A study of Bilirubin levels in patients with Liver Disease



58

Logarithm of Bilirubin Data



1. **Produce an interval estimate for the Population MEAN**
   log **bilirubin level**

2. **take anti-logs/exponentials of the resulting interval**

59

If Normality is questionable

a) Try to transform the data to approximate Normality
   - e.g. logarithms or square root

b) Non-Parametric technique
   - Bootstrap
   - CI for the population MEDIAN

60

The Bootstrap

a) Try to transform the data to approximate Normality
   - e.g. logarithms or square root

b) Non-Parametric technique
   - Bootstrap
   - CI for the population MEDIAN

61

61

Estimation via bootstrapping

- We can quantify the variability of sample statistics using theory eg the Central Limit Theorem, or by simulation via bootstrapping.

- The term **bootstrapping** comes from the phrase "pulling oneself up by one's bootstraps".

62

62

Bootstrapping scheme

- Take a bootstrap sample - a random sample taken with replacement from the original sample, of the same size as the original sample.
- Calculate the bootstrap statistic - a statistic such as mean, median, proportion, etc. computed on the bootstrap samples.
- Repeat steps (1) and (2) many times to create a bootstrap distribution - a distribution of bootstrap statistics.
- Calculate the bounds of the XX% confidence interval as the middle XX% of the bootstrap distribution.

63

63

Bootstrapping in R

```
# install.packages("infer")
library(infer)
```

64

64

Generate bootstrap means

```r
boot <- train.df %>%
  specify(response = Improvement) %>%
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "mean")

percentile_ci <- get_ci(boot)
round(percentile_ci,2)
```

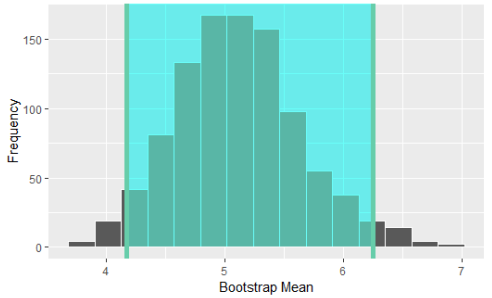
```
# A tibble: 1 x 2
  `2.5%` `97.5%`
   <dbl>   <dbl>
1   4.17    6.25
```

65

65

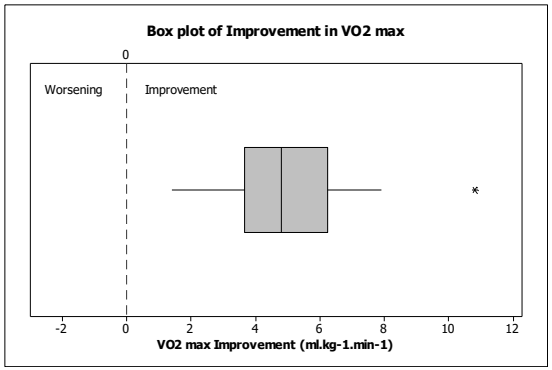
Plot the (empirical) sampling distribution

```r
boot %>% visualize(endpoints = percentile_ci, direction = "between")
+
                  xlab("Bootstrap Mean") + ylab("Frequency")
```



66

66



Variable	N	Mean	StDev
VO2 Improvement	18	5.11111	2.25829

67

67

Compare the two 95% Confidence Intervals

```
        One Sample t-test

data:  .
t = 9.6022, df = 17, p-value = 2.798e-08
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 3.988090 6.234132
```

```
# A tibble: 1 x 2
  `2.5%` `97.5%`
   <dbl>   <dbl>
1   4.17    6.25
```

68

68

Generate bootstrap medians

```r
boot <- train.df %>%
  specify(response = Improvement) %>%
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "mean")

percentile_ci <- get_ci(boot)
round(percentile_ci,2)
```

69

Generate bootstrap medians

```r
boot.median <- train.df %>%
  specify(response = Improvement) %>%
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "median")

percentile_ci_median <- get_ci(boot.median)
round(percentile_ci_median,2)
```

```
# A tibble: 1 x 2
  `2.5%` `97.5%`
   <dbl>   <dbl>
1    4.1    6.05
```

70

Celtic Study

- Based on the data provided the sample mean improvement was 5.11 mL/kg/min. We are 95% confident that the typical improvement in VO2 max is likely to be between 4 and 6 mL/kg/min.
- Given that the typical VO2 max at the start of this study was 67.66, the estimated typical improvement is approximately 7% (i.e. 5.11/67.66 expressed as percentage is 0.07*100).
- How would you translate this ?

71

Celtic Study

- Does this mean that each player will improve by 5.11 units ?

72

Pick a parameter of interest ….

1. *Estimate it using an (unbiased) estimator*
2. *Calculate its corresponding standard error*;
3. *Calculate the corresponding (1-α)100% CI*;
4. Check the terms and conditions
5. Report the conclusions of the analysis.

73

Effect of increasing the confidence level

90% C.I. for μ, $\quad \bar{x} \pm 1.65 \frac{s}{\sqrt{n}}$

95% C.I. for μ, $\quad \bar{x} \pm 1.96 \frac{s}{\sqrt{n}}$

99% C.I. for μ, $\quad \bar{x} \pm 2.58 \frac{s}{\sqrt{n}}$

74

Theorem 9.2

If $\bar{x}$ is used as an estimate of μ, we can be $100(1-\alpha)\%$ confident that the error will not exceed a specified amount e when the sample size is

$$n = \left(\frac{z_{\alpha/2}\sigma}{e}\right)^2.$$

Very useful for sample size calculations

75