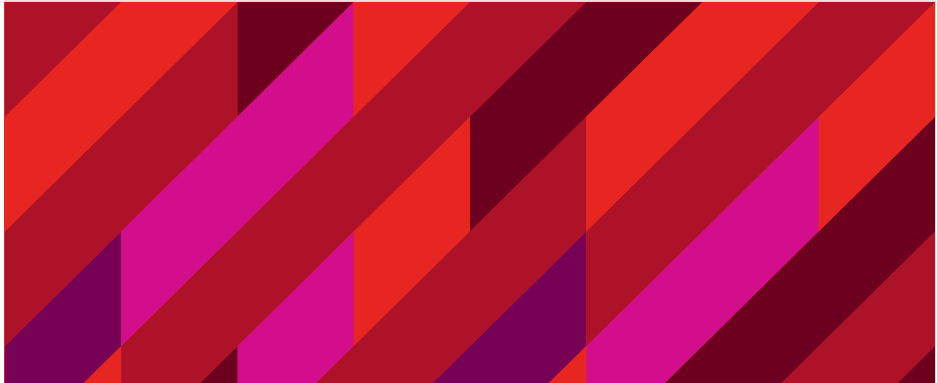




MACQUARIE
University
SYDNEY • AUSTRALIA

STAT2170 and STAT6180: Applied Statistics

Week 2: Modified two-sample t-test and test validation; Sample Size and Power





MACQUARIE
University
SYDNEY • AUSTRALIA

Two independent samples - two sample t test

- ▶ Twenty identical laboratory rats are randomly divided into two groups. Each group is fed a different diet. After a certain time period the weight gain (in g) of each rat was taken.

Diet1	73	121	110	81	105	89	128	83	112		
Diet2	137	97	152	103	103	138	135	146	129	133	119

- ▶ Is there any difference in weight between the two diets?

Data summary

Diet 1	$n_1 = 9$	$\bar{x}_1 = 100.22$	$s_1 = 19.34$
Diet 2	$n_2 = 11$	$\bar{x}_2 = 126.55$	$s_2 = 18.51$

- ▶ Define μ_i = true mean weight gain for rats using Diet i .
- ▶ In this case we don't care what values μ_1 and μ_2 take, but whether the two means are equal or different.
- ▶ Carry out a two sample (unpaired) t-test

$$H_0 : \mu_1 = \mu_2; \quad H_1 : \mu_1 \neq \mu_2$$

- ▶ Two options:
 - ▶ $\mu_1 = \mu_2$ and the observed difference between in weight gain between \bar{x}_1 and \bar{x}_2 can be explained by random variation; or
 - ▶ The sample means for Diets 1 and 2 are so far apart that we can conclude that we have evidence that the two populations means are different, ie, $\mu_1 \neq \mu_2$.
- ▶ Define σ_i = true standard deviation of weight gain measurements for rats using Diet i .
- ▶ The t distribution requires a single estimate of variance. We need to assume $\sigma_1^2 = \sigma_2^2$ (often quite valid), which can be checked graphically using a boxplot.
- ▶ Firstly we need to estimate the pooled (common) variance:

$$\hat{\sigma}^2 = s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

with $n_1 + n_2 - 2$ degrees of freedom.

- The test statistic is:

$$t_{obs} = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

- Compare t_{obs} with $t_{n_1+n_2-2}$, and obtain the P-Value for the two tailed test

$$\text{P-Value} = P(|t_{n_1+n_2-2}| \geq |t_{obs}|).$$

- Then, compare P-Value with α , etc, by following the hypothesis test procedure described last week.
- For the current example, the pooled sample variance/sd is

$$\begin{aligned} s_p^2 &= \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \\ &= \frac{(9 - 1) \times 19.34^2 + (11 - 1) \times 18.51^2}{9 + 11 - 2} \\ &= 356.5712682 \\ s_p &= \sqrt{s_p^2} \approx 18.8830948 \end{aligned}$$

- ▶ Recall the Hypotheses, $H_0 : \mu_1 = \mu_2$, $H_1 : \mu_1 \neq \mu_2$
 - ▶ Use $\alpha = 0.05$ as usual.

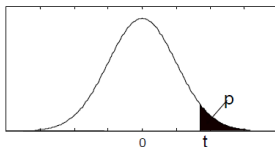
- ▶ **Test statistic**

$$\begin{aligned}t_{obs} &= \frac{\bar{X}_1 - \bar{X}_2}{s_p \sqrt{1/n_1 + 1/n_2}} \\&= \frac{100.22 - 126.55}{18.8830948 \times \sqrt{\frac{1}{9} + \frac{1}{11}}} \\&= \frac{-26.33}{8.487} \\&= -3.102\end{aligned}$$

- ▶ Degrees of freedom (df) = $n_1 + n_2 - 2 = 9 + 11 - 2 = 18$
- ▶ Compute P-Value based on t_{obs} under the $t_{n_1+n_2-2} = t_{18}$ -distribution.

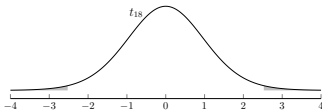
Critical values t-distribution

$0.001 < P(t_{18} > | -3.102 |) < 0.005$
 $0.001 < P(t_{18} > 3.102) < 0.005$
 $2 * 0.001 < 2 * P(t_{18} > 3.102) < 2 * 0.005$
 $0.002 < P\text{-value} < 0.01$
 $P\text{-value} < 0.01$ (i.e. less than 0.05)



df	0.250	0.200	0.150	0.100	0.050	0.025	0.020	0.010	0.005	0.001
1	1.0000	1.3764	1.9626	3.0777	6.3138	12.7062	15.8945	31.8205	63.6567	318.3088
2	0.8165	1.0607	1.3862	1.8856	2.9200	4.3027	4.8487	6.9646	9.9248	22.3271
3	0.7649	0.9785	1.2498	1.6377	2.3534	3.1824	3.4819	4.5407	5.8409	10.2145
4	0.7407	0.9410	1.1896	1.5332	2.1318	2.7764	2.9985	3.7469	4.6041	7.1732
5	0.7267	0.9195	1.1558	1.4759	2.0150	2.5706	2.7565	3.3649	4.0321	5.8934
6	0.7176	0.9057	1.1342	1.4398	1.9432	2.4469	2.6122	3.1427	3.7074	5.2076
7	0.7111	0.8960	1.1192	1.4149	1.8946	2.3646	2.5168	2.9980	3.4995	4.7853
8	0.7064	0.8889	1.1081	1.3968	1.8595	2.3060	2.4490	2.8965	3.3554	4.5008
9	0.7027	0.8834	1.0997	1.3830	1.8331	2.2622	2.3984	2.8214	3.2498	4.2968
10	0.6998	0.8791	1.0931	1.3722	1.8125	2.2281	2.3593	2.7638	3.1693	4.1437
11	0.6974	0.8755	1.0877	1.3634	1.7959	2.2010	2.3281	2.7181	3.1058	4.0247
12	0.6955	0.8726	1.0832	1.3562	1.7823	2.1788	2.3027	2.6810	3.0545	3.9296
13	0.6938	0.8702	1.0795	1.3502	1.7709	2.1604	2.2816	2.6503	3.0123	3.8520
14	0.6924	0.8681	1.0763	1.3450	1.7613	2.1448	2.2638	2.6245	2.9768	3.7874
15	0.6912	0.8662	1.0735	1.3406	1.7531	2.1314	2.2485	2.6025	2.9467	3.7328
16	0.6901	0.8647	1.0711	1.3368	1.7459	2.1199	2.2354	2.5835	3.102 lies in this range	
17	0.6892	0.8633	1.0690	1.3334	1.7396	2.1098	2.2238	2.5669		
18	0.6884	0.8620	1.0672	1.3304	1.7341	2.1009	2.2137	2.5524	2.8784	3.6105

$$\begin{aligned}\text{P-Value} &= 2P(t_{n_1+n_2-2} > |t_{obs}|) \\ &= 2P(t_{18} > |-3.102|) \\ &= 0.0061523 \text{ (exact)} \\ &< 2 \times 0.005 = 0.01 \text{ (using table)} \\ &< 0.05\end{aligned}$$



- ▶ We can conclude at the 5% level of significance that the two diets do not have an equal effect on weight gain of the rats. In fact, Diet 2 resulted in significantly higher weight gain.
- ▶ Note: The test conducted could be two tailed or one tailed depending on the research question being asked. If Diet 1 was the *standard* diet, we may want to test if the new diet (diet 2) results in **higher** weight gain. This would imply $H_1 : \mu_1 < \mu_2$ as the appropriate alternative hypothesis. In this case, the P-Value is less than 0.005, half of the 0.01 shown above.

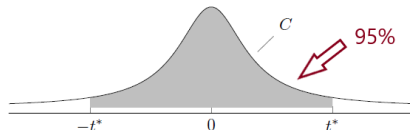


Confidence interval for $\mu_1 - \mu_2$

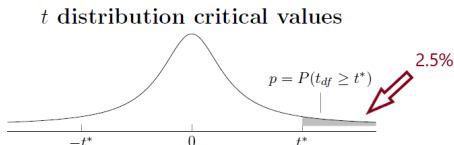
- ▶ Remember all confidence intervals of a parameter are in the form:
estimate \pm critical value \times estimated s.e.(estimate)
- ▶ In this case:
 - ▶ Parameter = $\mu_1 - \mu_2$
 - ▶ Estimate = $\bar{x}_1 - \bar{x}_2 = 100.22 - 126.55 = -26.33$
 - ▶ Critical Value = $t_{n_1+n_2-2, 1-\alpha/2}^{\alpha=0.05} = t_{18, 0.975} = 2.101$
 - ▶ S.E. Estimate = $s_p \sqrt{1/n_1 + 1/n_2} = 8.487$

Obtaining critical value - table option 1

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
	.50	.60	.70	0.80	.90	.95	.96	.98	.99	.995	.998	.999
Probability C												



Obtaining critical value - table option 2



Key: Table entry for p and C is the critical value t^* with probability p lying to its right and probability C lying between $-t^*$ and t^* .

		Upper tail probability p											
df	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005	
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82	63.66	127.3	318.3	636.6	
2	0.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.09	22.33	31.60	
3	0.765	0.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.21	12.92	
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	1.730	2.093	2.207	2.540	2.861	3.174	3.583	3.890	

For the example:

- ▶ 95% CI for $\mu_1 - \mu_2$ is $(\bar{x}_1 - \bar{x}_2) \pm t_{n_1+n_2-2, \alpha/2} s_p \sqrt{1/n_1 + 1/n_2}$
 $-26.33 \pm 2.101 \times 8.487 = -26.33 \pm 17.831 = (-44.161, -8.499)$
- ▶ We are 95% confidence that the interval $(-44.16, -8.5)$ includes the true difference between the means

- ▶ Confidence interval can be used for two tailed tests of significance.
 - ▶ $H_0 : \mu_1 = \mu_2$; and $H_1 : \mu_1 \neq \mu_2$ is equivalent to
 - ▶ $H_0 : \mu_1 - \mu_2 = 0$; and $H_1 : \mu_1 - \mu_2 \neq 0$
- ▶ Outcome of the test at level α can be determined by checking if zero lies inside the $100(1 - \alpha)\%$ confidence interval (CI).
 - ▶ Reject $H_0 : \mu_1 - \mu_2 = 0$ at level α if zero lies **outside** $100(1 - \alpha)\%$ CI for $\mu_1 - \mu_2$
 - ▶ Don't Reject $H_0 : \mu_1 - \mu_2 = 0$ at level α if zero lies **inside** $100(1 - \alpha)\%$ CI for $\mu_1 - \mu_2$
- ▶ For our example.
 - ▶ 95% CI for $\mu_1 - \mu_2 = (-44.16, -8.5)$
 - ▶ Zero lies outside the 95% CI, meaning that we would reject H_0 at the 5% level of significance ($\alpha = 0.05$).



MACQUARIE
University
SYDNEY • AUSTRALIA

Assumptions

Assumptions for the two sample t -test

- ▶ The two samples are independent (ie, not related in any way).
 - ▶ This can't really be checked unless we have been involved in the collection of the data.
- ▶ Observations in each sample are independent of each other (usually can't be checked but rely on the design of study in data collection to ensure such independence).
- ▶ The data come from a Normal or an approximate Normal distribution.
 - ▶ Later in this lecture we can check this using normal quantile plots.
- ▶ The standard deviations (variances) for both populations are the same, ($\sigma_1 = \sigma_2$)
 - ▶ Check this using **boxplot**
 - ▶ Check how close s_1 and s_2 . In our example, $s_1 = 19.34$ and $s_2 = 18.51$.
 - ▶ There are tests available for unequal variances.



MACQUARIE
University
SYDNEY • AUSTRALIA

Using R: Two sample t -test

Using R: Load the Diet data

- ▶ Use the RStudio *Tools* → *Import Dataset* menu *OR*
- ▶ Use the command line
- ▶ Note below, there are two variables
 - ▶ **gain**: weight gain for each rat
 - ▶ **diet**: indicator which diet for each rat.

```
diet.dat = read.table('diet.dat', header = TRUE)  
head(diet.dat)
```

```
#   gain  diet  
# 1   73 Diet1  
# 2  121 Diet1  
# 3  110 Diet1  
# 4   81 Diet1  
# 5  105 Diet1  
# 6   89 Diet1
```

```
t.test(gain ~ diet, var.equal = TRUE, data = diet.dat)
```

```
#  
# Two Sample t-test  
#  
# data: gain by diet  
# t = -3.1015, df = 18, p-value = 0.006159  
# alternative hypothesis: true difference in means between group Diet1 and group  
# 95 percent confidence interval:  
# -44.154430 -8.492035  
# sample estimates:  
# mean in group Diet1 mean in group Diet2  
# 100.2222 126.5455
```

Using R: Alt. method, two data objects

```
diet1 = c(73, 121, 110, 81, 105, 89, 128, 83, 112)
diet2 = c(137, 97, 152, 103, 103, 138, 135, 146, 129,
          133, 119)
head(diet1)
```

```
# [1] 73 121 110 81 105 89
```

```
head(diet2)
```

```
# [1] 137 97 152 103 103 138
```

```
t.test(diet1, diet2, var.equal = TRUE)
```

```
#  
#   Two Sample t-test  
#  
# data:  diet1 and diet2  
# t = -3.1015, df = 18, p-value = 0.006159  
# alternative hypothesis: true difference in means is not equal to 0  
# 95 percent confidence interval:  
#  -44.154430  -8.492035  
# sample estimates:  
# mean of x mean of y  
#  100.2222  126.5455
```



MACQUARIE
University
SYDNEY • AUSTRALIA

Paired data and paired t test; Confidence interval

- ▶ Suppose a different experiment is conducted where similar individuals are **paired**. E.g.
 - ▶ Pair 1: Two fattest rats
 - ▶ Pair 2: Next two fattest rats
 - ▶ \vdots
 - ▶ Pair n : Thinnest two rats.
- ▶ In this way, we get n pairs of rats.
 - ▶ Their initial weight is similar by design.
 - ▶ Each rat in the pair given different diet (assigned at random).

- ▶ We assume the two rats in each pair are the same (having similar “initial fatness”/weight), and randomly allocate one to Diet 1 and the other one to Diet 2.
- ▶ We assume the only difference in weight after being on Diet 1 or Diet 2 between the two rats in each pair is due to the treatment.
- ▶ There may or may not be differences between the pairs. Those differences can be great or small. However, it doesn't matter as the differences between the pairs are excluded from the analysis.
- ▶ Only the difference within each pair is analysed. This is done by taking the difference for each pair and testing/analysing the differences from all n pairs.

In order to investigate whether the time of neutering of male dogs had any effect on their rate of growth, Crenshaw and Carter (1995) neutered dogs at 2 months and 7 months of age and measured their weight at 18 months of age. Two male dogs were randomly selected from each available litter (with similar weight). One was neutered at 2 months, the other was neutered at 7 months. All the dogs were kept under identical conditions during the 18 months of the trial. (paired by litter)

Litter	Two.months	Seven.months	difference
1	19.0	18.9	-0.1
2	23.2	23.5	0.3
3	16.1	16.7	0.6
4	30.3	30.3	0.0
5	27.1	27.8	0.7
6	22.3	22.4	0.1
7	18.1	18.2	0.1
8	27.8	28.1	0.3
9	34.2	34.1	-0.1
10	17.3	17.7	0.4

Data summary

Differences	$n_d = 10$	$s_d = 0.2791$	$\bar{d} = 0.23$
-------------	------------	----------------	------------------

Test if mean difference is zero (no difference).



- ▶ One sample t test on the differences (known as paired t test).

$$H_0 : \mu_d = 0; \quad H_1 : \mu_d \neq 0$$

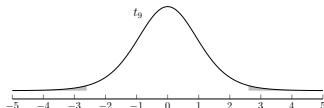
- ▶ Test statistic

$$\begin{aligned} t_{obs} &= \frac{\bar{d}}{s_d/\sqrt{n}} \\ &= \frac{0.23}{0.2791/\sqrt{10}} \\ &= 2.605962 \end{aligned}$$

- ▶ Compare t_{obs} against a t_9 distribution ($df = 10 - 1 = 9$)

Compute Paired t-test P-Value

$$\begin{aligned}\text{P-Value} &= 2P(t_{n-1} > |t|) \\ &= 2P(t_9 > |2.606|) \\ &= 2 \times 0.01436911 \text{ (exact)} \\ &= 0.02845682 \\ &< 2 \times 0.02 = 0.04 \text{ (using tables)} \\ &< 0.05\end{aligned}$$



- ▶ Therefore, reject H_0 at the 5% level of significance as the P-Value is less than 0.05.
 - ▶ Conclude that there is evidence of a significant change in average weight at 18 months of age for the dogs neutered at two months and seven months.
- ▶ Note that the test could have been carried out as a one tailed test if appropriate.

Assumptions for the paired t test and relevant confidence intervals:

- ▶ We assume that the differences are approximately Normally distributed. (need to check normality)
- ▶ We are making absolutely no assumptions about the original observations. In fact, if they are not Normally distributed, it doesn't matter, as long as the differences are Normally distributed.
- ▶ We are also assuming all the differences are independent.

- ▶ Remember all confidence intervals of a parameter are in the form:
estimate \pm critical value \times estimated s.e.(estimate)
- ▶ In this case:
 - ▶ Parameter = μ_d
 - ▶ Estimate = $\bar{d} = 0.23$
 - ▶ Critical Value = $t_{n_d-1, 1-\alpha/2} = t_{9, 0.975} = 2.262$
 - ▶ S.E. Estimate = $\frac{s_d}{\sqrt{n_d}} = 0.088$
- ▶ 95% CI for μ_d is $\bar{d} \pm t_{n_d-1, \alpha/2} \frac{s_d}{\sqrt{n_d}}$
 $0.23 \pm 2.262 \times 0.088 = 0.23 \pm 0.199 = (0.031, 0.429)$
- ▶ We are 95% confidence that the interval (0.031, 0.429) includes the true difference between the means

Using R for paired t-test, read data

```
dogs = read.table("dogs.dat", header = TRUE)  
head(dogs)
```

#	Two.months	seven.months	difference
# 1	19.0	18.9	-0.1
# 2	23.2	23.5	0.3
# 3	16.1	16.7	0.6
# 4	30.3	30.3	0.0
# 5	27.1	27.8	0.7
# 6	22.3	22.4	0.1

```
t.test(dogs$seven.months, dogs$Two.months, paired = TRUE)
```

```
#  
#   Paired t-test  
#  
# data:  dogs$seven.months and dogs$Two.months  
# t = 2.6061, df = 9, p-value = 0.02845  
# alternative hypothesis: true difference in means is not equal to 0  
# 95 percent confidence interval:  
#  0.03035404 0.42964596  
# sample estimates:  
# mean of the differences  
#                      0.23
```




MACQUARIE
University
SYDNEY • AUSTRALIA

Paired vs Two sample t-tests

- ▶ To do a paired t-test, two groups of data collected must be **paired** (in pairs by nature or construction).
- ▶ For two independent groups/samples, a two sample t-test is appropriate.
- ▶ Examples of paired data
 - ▶ before and after
 - ▶ left and right side of a person/animal
 - ▶ top and bottom leaves of a plant
 - ▶ two animals in cage/two plants in a pot
 - ▶ two individuals at the same age/weight
 - ▶ two twins

CAUTION: Using inappropriate test

- ▶ Suppose a two sample test was conducted on the dog data (ignoring the pairing by litter).

```
t.test(dogs$seven.months, dogs$Two.months, var.equal = TRUE)
```

```
#  
# Two Sample t-test  
#  
# data: dogs$seven.months and dogs$Two.months  
# t = 0.084666, df = 18, p-value = 0.9335  
# alternative hypothesis: true difference in means is not equal to 0  
# 95 percent confidence interval:  
# -5.477247 5.937247  
# sample estimates:  
# mean of x mean of y  
# 23.77 23.54
```

- ▶ Contrast the P-values and outcomes
 - ▶ Paired test: P-value < 0.05 , so we reject H_0
 - ▶ Two sample test: P-value $= 0.93 \gg 0.05$, we *don't* reject H_0 .
- ▶ Clearly, can be misleading and lead to **incorrect** conclusions if the wrong test is used!
- ▶ Suppose now the data is not paired and the samples are independent.
 - ▶ *BUT* the sample standard deviations are very different
 - ▶ This violates the assumption of the two sample *t*-test earlier.
 - ▶ We need a separate test in this scenario.



MACQUARIE
University
SYDNEY • AUSTRALIA

Modified Two sample t-test

Modified Two sample t-test

Two sample t-test with unequal variances.



Example: Obtain the concentration of a chemical by two different methods. Does the new, faster method give the same result as the standard?

Standard	25	24	25	26					
New	23	18	22	28	17	25	19	16	

Data summary

Standard	$n_1 = 4$	$\bar{x}_1 = 25$	$s_1 = 0.82$
New	$n_2 = 8$	$\bar{x}_2 = 21$	$s_2 = 4.21$

- Notice the big discrepancy between s_1 and s_2 . We cannot use a test that assumes that σ_1 and σ_2 are equal.

Two sample t-test not assuming equal variances



- ▶ We wish to test $H_0 : \mu_1 = \mu_2$; $H_1 : \mu_1 \neq \mu_2$
 - ▶ Ratio of standard deviations is large $4.21/0.82 = 5.1341463$
 - ▶ Even though the sample sizes are small, ratio is far too large to assume equal variances.
 - ▶ We cannot pool the variances into a single estimate.
- ▶ There is no exact test in this situation. All we can do is carry out an approximate test, based the following test statistic:

$$t_{obs} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

- ▶ The test follows the same format as the two sample t-test assuming equal variances except that the degrees of freedom have to be adjusted for this (approximate) t test.
 - ▶ The degrees of freedom (df) in this case is given by,

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1-1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2-1}}.$$

- ▶ known as Welch-Satterthwaite modification.
- ▶ **Note:** The degrees of freedom (df) are often not an integer and if computing by hand we round down to the closest integer, or just get R to do the test.

```
chem.dat = read.table("chem.dat", header = TRUE)  
head(chem.dat)
```

```
#   concen  method  
# 1      25 Standard  
# 2      24 Standard  
# 3      25 Standard  
# 4      26 Standard  
# 5      23      New  
# 6      18      New
```

- ▶ Note again, there are two variables per observation in this data:
 - ▶ **concen**: Numeric value of chemical concentrations.
 - ▶ **method**: Indicator for the method used.

```
t.test(concen ~ method, data = chem.dat)
```

```
#  
#   Welch Two Sample t-test  
#  
# data:  concen by method  
# t = -2.5923, df = 7.9878, p-value = 0.03204  
# alternative hypothesis: true difference in means between group New and group  
# 95 percent confidence interval:  
#  -7.5591856 -0.4408144  
# sample estimates:  
#      mean in group New mean in group Standard  
#                   21                   25
```

► Note: Unequal variances are the default option (see ? t.test)

Have a go at writing up your own notes as the unit progresses. E.g.:

100 (1- α) % CI for Parameter: Estimate \pm critical value \times s.e. (estimate)

one sample
 σ known μ :

$$\bar{x} \pm z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}$$

one sample
 σ unknown μ :

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

two samples
paired
(one sample of
differences)
 σ_d unknown μ_d :

$$\bar{d} \pm t_{n_d-1, \alpha/2} \times \frac{s_d}{\sqrt{n_d}}$$

two samples
 $\sigma = \sigma_1 = \sigma_2$ unknown
equal var $\mu_1 - \mu_2$:

$$\bar{x}_1 - \bar{x}_2 \pm t_{n_1+n_2-2, \alpha/2} \times s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Another example

ONE SAMPLE t-test

$$H_0: \mu = \mu_0$$

$$H_1: \mu \neq \mu_0$$

$$t_{\text{obs}} = \frac{\bar{x} - \mu_0}{\left(\frac{s}{\sqrt{n}}\right)}$$

↓ compare
to

$$t_{n-1}$$

Two SAMPLE t-test

$$H_0: \mu_1 = \mu_2$$

$$H_1: \mu_1 \neq \mu_2$$

$$t_{\text{obs}} = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

↓ compare
to

$$t_{n_1+n_2-2}$$

Paired t-test

$$H_0: \mu_d = 0$$

$$H_1: \mu_d \neq 0$$

$$t_{\text{obs}} = \frac{\bar{d}}{\left(\frac{s_d}{\sqrt{n}}\right)}$$

↓ compare
to

$$t_{n_d-1}$$

Welch Two SAMPLE

t-test
(not assuming equal variances)

$$H_0: \mu_1 = \mu_2 ; H_1: \mu_1 \neq \mu_2$$

$$t_{\text{obs}} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n} + \frac{s_2^2}{n}}}$$

↓ compare
to

t_v ; where

$$v = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1-1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2-1}}$$

< compare p-value with level of significance. Write a statistical and contextual conclusion. (This step is the same regardless of statistical test.) >

Add to these notes after the lecture by including the assumptions of each test and how they can be checked.



MACQUARIE
University
SYDNEY • AUSTRALIA

Assessing Normality

- ▶ Whenever we assume a Normal distribution, we should determine whether our assumption is acceptable. Up to now:
 - ▶ Obtain a histogram, and look for
 - ▶ a symmetric distribution and
 - ▶ an approximate *bell-shape*.
- ▶ Graphing will give you an idea of the shape, skewness, etc, but it doesn't specifically look at normality.
- ▶ There are significance tests for testing for a Normal distribution:
 - ▶ H_0 : Data come from a Normal population
 - ▶ but they require a reasonably large amount of data before we can make a conclusion with any confidence.

- ▶ For small data sets, significance test for Normality may be inconclusive (ie, non-significant) even for the most obvious of data sets (randomly selected from a normal distribution).
- ▶ However, we can get an idea of Normality by obtaining and then checking out the Normal Probability (Normality) Plot of the data.
- ▶ This will give us a subjective measure of how closely our data match a Normal distribution.

- ▶ Also known as a normal QQ plot, steps require to create are below.

1. Calculate the Normal scores:

- ▶ These are the values along the axis that split the bell curve up into equal sized areas (surrogates for values behaving perfectly on the standard normal bell curve, $Z \sim \mathcal{N}(0, 1)$).

2. Match up the normal scores with the sample data:

- ▶ Pair up the ordered sample data with the ordered normal scores.
i.e. smallest sample observed paired with the smallest normal score.

3. Plot the pairs on a scatterplot:

- ▶ If the data on the scatterplot are close to a straight line, this indicates that our data are very close to a Normal distribution (**Note:** a perfect straight line is very suspicious of fake data)

- ▶ We don't expect the line to be exactly straight. In fact if the plot gave a perfectly straight line, it would indicate that there is no random variation in the data.
- ▶ We expect random variation in the data, so, if the data come from a Normal distribution, we would expect random variation about a straight line.
- ▶ The further the plot is from a straight line, the less confident we are about if our data are Normally distributed.
- ▶ We expect there to be no consistent patterns in the deviation from the line.
- ▶ The calculations are time consuming, so get R to do it all.

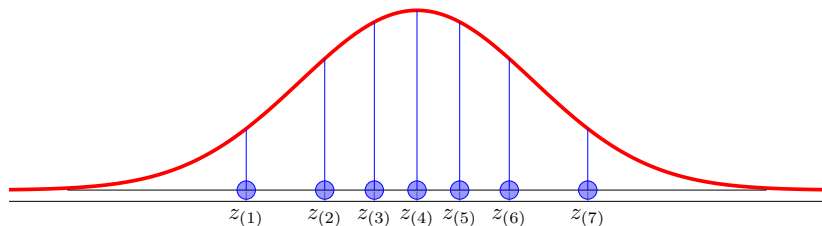
Example: Normal Quantile plot

X	10.37	8.7	8.52	9.7	9.51	11.16	10.76
---	-------	-----	------	-----	------	-------	-------

Is it reasonable to assume this data come from a Normal distribution?

Calculation of the Normal scores

- ▶ There are 7 observations, so the area under the standard Normal curve is divided into 8 equal areas.
- ▶ The normal scores are the values that divide the area into equal parts.
 - ▶ 7 ordered z values, $z_{(1)}, z_{(2)}, \dots, z_{(7)}$.
- ▶ The $z_{(i)}$ values are not equally spaced. It is the **areas** that are equal.
- ▶ **Note:** the actual data has no effect on the values of the N-scores. It is only the number of observations that determines the partitioning of the Z distribution (see next slide).



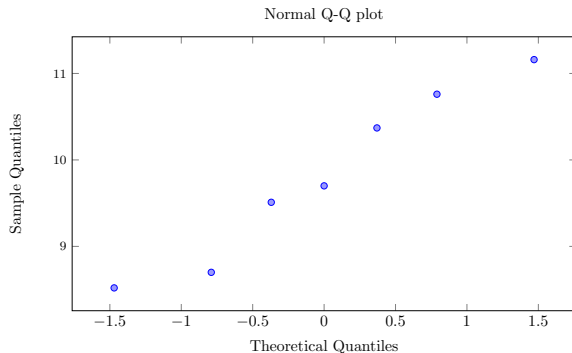
- ▶ Match up the $z_{(i)}$ values with the data
 - ▶ $z_{(1)}$ paired with smallest observed data point.
 - ▶ $z_{(2)}$ paired with second smallest observed data point.
 - ▶ \vdots
 - ▶ $z_{(7)}$ paired with largest observed data point.

Match the normal scores against the data

Before Sorting		After sorting	
data	normal score	data	normal score
10.37	0.37	8.52	-1.47
8.7	-0.79	8.7	-0.79
8.52	-1.47	9.51	-0.37
9.7	0.00	9.7	0.00
9.51	-0.37	10.37	0.37
11.16	1.47	10.76	0.79
10.76	0.79	11.16	1.47

Plot the normal scores against the data

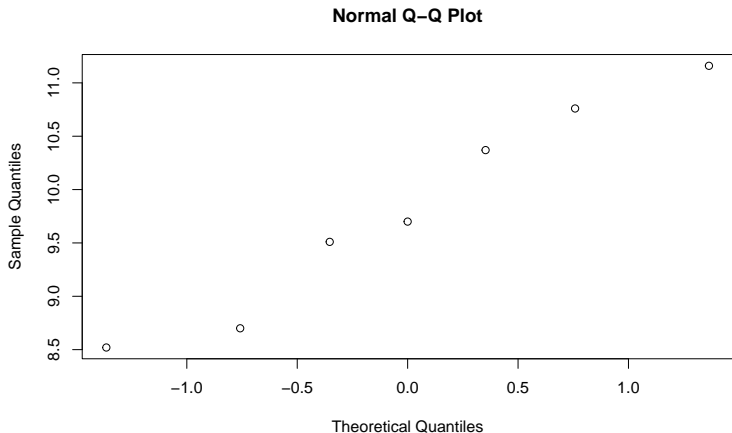
- ▶ Both ordered.



- ▶ Determine the extent to which the points follow a straight line.

Using R: Create a Normal QQ Plot

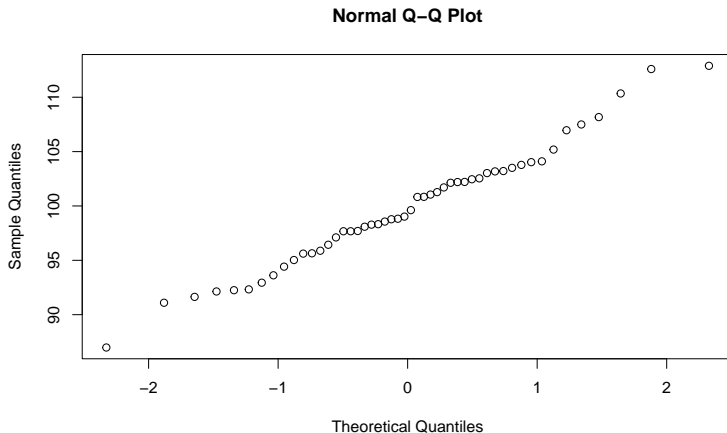
```
x = c(10.37, 8.70, 8.52, 9.70, 9.51, 11.16, 10.76)  
qqnorm(x)
```



Normal QQ Plot: Normal($20, 5^2$) Distribution



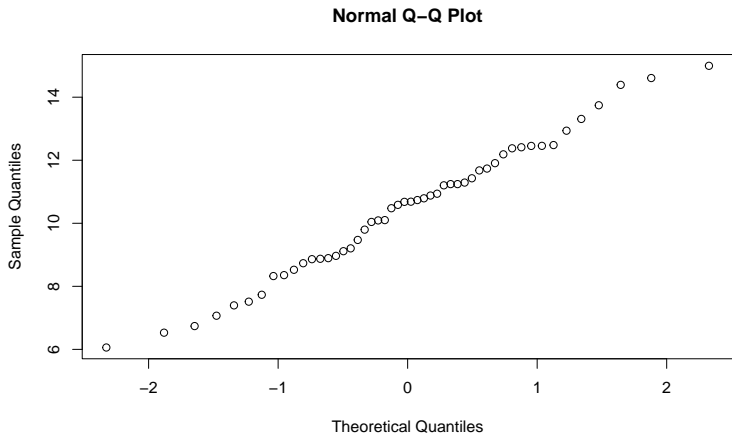
```
x = rnorm(50, mean = 100, sd = 5)  
qqnorm(x)
```



Normal QQ Plot: Normal($10, 2^2$) Distribution

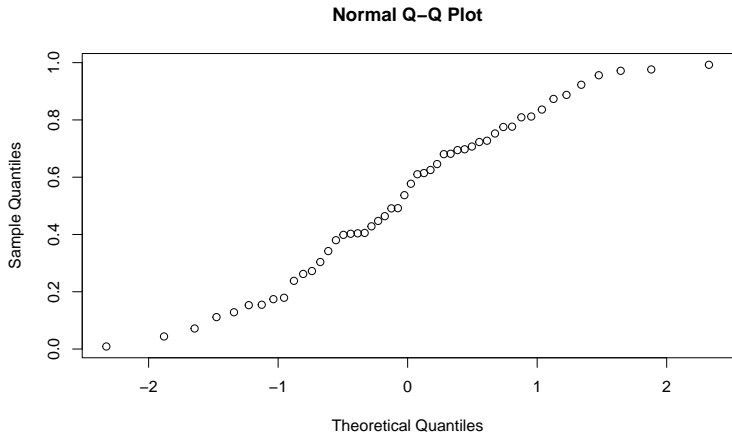


```
x = rnorm(50, mean = 10, sd = 2)  
qqnorm(x)
```



Normal QQ Plot: Uniform Distribution

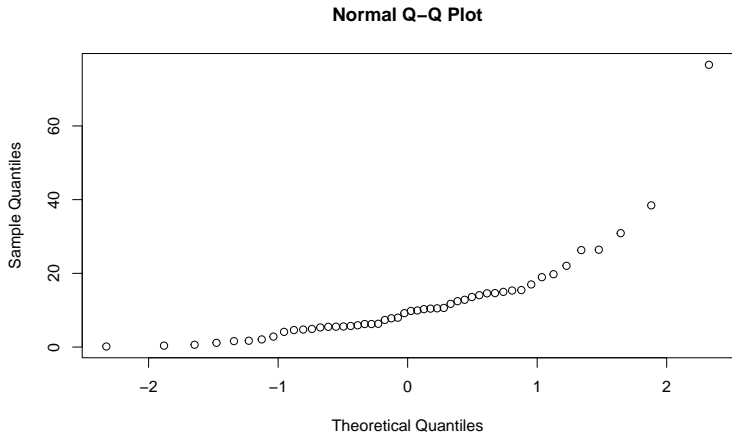
```
u = runif(50)  
qqnorm(u)
```



Normal QQ Plot: Gamma Distribution



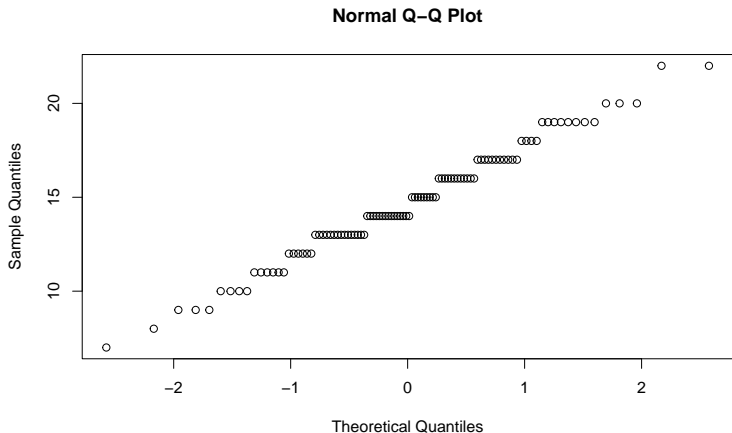
```
g = rgamma(50, shape = 1, scale = 10)  
qqnorm(g)
```



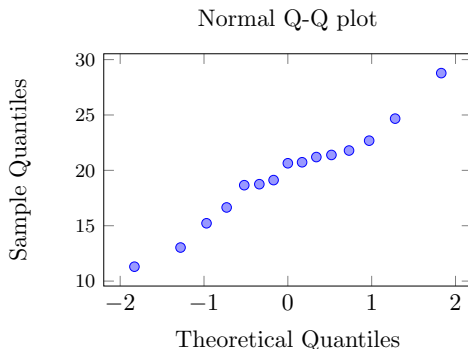
Normal QQ Plot: Binomial(100, 0.3) Distribution

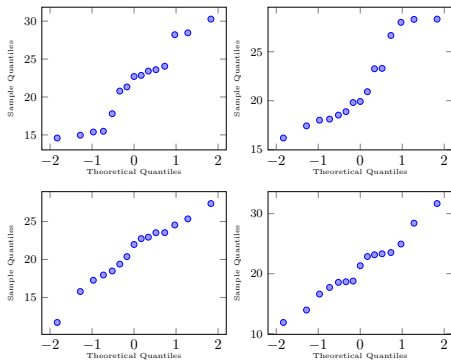


```
b = rbinom(50, n = 100, prob = 0.3)
qqnorm(b)
```



- ▶ If the sample size is reasonably small, we can expect a fair amount of random variation of the data from a straight line.
 - ▶ E.g.: A sample of size $n = 15$ from $\mathcal{N}(20, 5^2)$

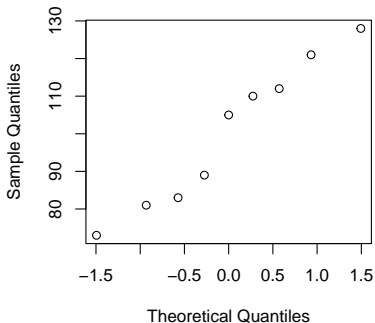




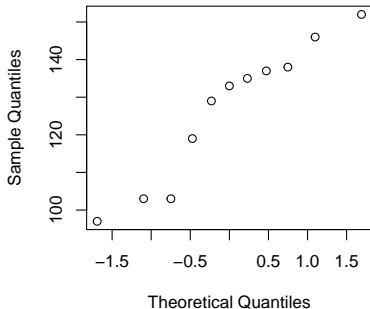
- ▶ When checking the Normality of the data for two sample t-tests we need to obtain a separate Normality plot for each sample. That is, we require:
 - ▶ Data from sample 1 close to a normal distribution
 - ▶ Data from sample 2 close to a normal distribution
- ▶ We may assume they have the same variance, but they could have different means, hence different Normal distributions.
 - ▶ So we must create separate Normal QQ plots for each sample.

```
diet1 = gain[diet == "Diet1"]  
diet2 = gain[diet == "Diet2"]  
qqnorm(diet1, main = "Normal Q-Q plot for Diet 1")  
qqnorm(diet2, main = "Normal Q-Q plot for Diet 2")
```

Normal Q-Q plot for Diet 1



Normal Q-Q plot for Diet 2





MACQUARIE
University
SYDNEY • AUSTRALIA

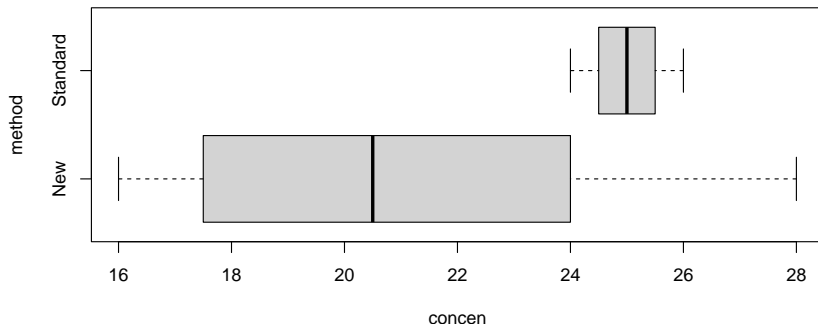
Checking equal variances

Checking $\sigma_1 = \sigma_2$ (equal variances)

- ▶ **Rule of Thumb:** The ratio of the standard deviations, $s_{larger}/s_{smaller}$ should be less than 2.
- ▶ The larger the difference between the two sample standard deviations, the more stringent we should be.
 - ▶ For very large sample sizes, we should aim for a ratio less than 1.5.
 - ▶ For small sample sizes, we could allow the ratio to be larger, say up to 2.5.
- ▶ If the ratio of the sample standard deviations is too large, we can't pool the variances (since we can't assume $\sigma_1 = \sigma_2$). Therefore the two sample t -test assuming equal variances is **NOT** appropriate.
 - ▶ In this case, use the approximate two sample t -test instead.

- Clearly can see that the standard method has a much smaller spread than the new method \Rightarrow can't use two sample t-test with equal variances.

```
boxplot(concen ~ method, data = chem.dat, horizontal = TRUE)
```



Checking for equal variances - distribution of sample variance

We know that if the sample, X_1, X_2, \dots, X_n , is normally distributed then the sample mean is also normally distributed, that is, $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$.

Recall that

$$s^2 = \sum_{i=1}^n (X_i - \bar{X})^2 / (n - 1)$$

It can be shown that distribution of $(n - 1)s^2$ is χ_{n-1}^2 .

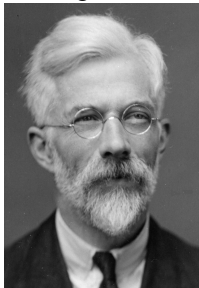
If we have two independent χ^2 variables with degrees of freedom ν_1 and ν_2 then,

$$\frac{\chi_{\nu_1}^2/\nu_1}{\chi_{\nu_2}^2/\nu_2} \sim F_{\nu_1, \nu_2}$$

has F-distribution with ν_1 and ν_2 degrees of freedom.

The ratio of two independent χ^2 values divided by their respective degrees of freedom is an F distribution with degrees of freedom given by (df numerator, df denominator).

- ▶ Ronald A. Fisher (1890-1962) was an English statistician and biologist.
 - ▶ Known as the father/creator for **modern statistical science**.
 - ▶ Also regarded as the greatest biologist since Charles Darwin.



- ▶ Distribution of ratio of χ^2 variables (appropriately rescaled) is named after Fisher.

- ▶ How does an F_{ν_1, ν_2} distribution behave?
 - ▶ ν_1 and ν_2 known as the degrees of freedom for the numerator and denominator.
- ▶ It is a ratio of χ^2 , so has the properties
 - ▶ Must be positive
 - ▶ Also right skew
- ▶ Mean value at $\nu_2/(\nu_2 - 2)$, i.e. close to 1.

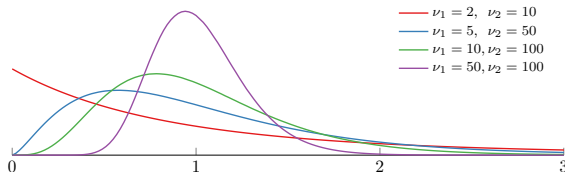


Figure 1: F_{ν_1, ν_2} -distribution examples

Critical Tail value: F -distribution

- ▶ The F_{ν_1, ν_2} -distribution is a *family* of distributions with
 - ▶ degrees of freedom ν_1, ν_2
 - ▶ **Caution:** Order is important, F_{ν_1, ν_2} is a **different** distribution to an F_{ν_2, ν_1} -distribution!
- ▶ For testing, large values of the observed statistic support H_1
- ▶ Find the **critical value** with α upper tail probability.

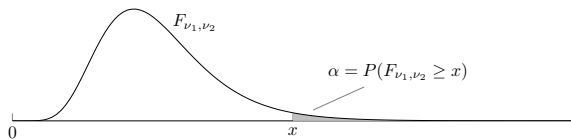


Figure 2: α upper tail probability

- ▶ $x = F_{\nu_1, \nu_2, 1-\alpha}$, the **critical value** to reject H_0 at level α

Checking $\sigma_1 = \sigma_2$ (equal variances)

- ▶ **Rigorous Test for Equal Variance:** The ratio of the sample variances, s_1^2/s_2^2 , follows a F -distribution with degrees of freedoms $n_1 - 1$ and $n_2 - 1$.
- ▶ If s_1^2/s_2^2 is larger than the 97.5% quantile or smaller than the 2.5% quantile of the F -distribution with degrees of freedoms $n_1 - 1$ and $n_2 - 1$, then at a level of 5% reject the hypothesis that the two populations share the same variance.

Test example - equal variances

- ▶ A fourth grade class has 7 girls and 10 boys. The children's heights are recorded on their 10th birthday as follows:

Girls	140.3	139.8	138.6	137.1	140	136.2	138.7			
Boys	135.3	137.0	136.0	139.7	136.5	137.2	138.8	139.6	140.0	137.7

- ▶ Do the data provide sufficient evidence to indicate a difference in the variability of heights for boys and girls?
- ▶ Assume boys height $X \sim N(\mu_X, \sigma_X^2)$ and the girls height $Y \sim N(\mu_Y, \sigma_Y^2)$
- ▶ The hypotheses to be tested are

$$H_0 : \sigma_X^2 = \sigma_Y^2 \text{ vs } H_1 : \sigma_X^2 \neq \sigma_Y^2$$

Test example - equal variances

```
ratio = var(fgrade$boys) / var(fgrade$girls)
```

```
# [1] 1.161809
```

```
m <- length(fgrade$boys)
n <- length(fgrade$girls)
pval <- 2* min (pf(ratio, m-1, n-1), 1 - pf(ratio, m-1, n-1))
```

```
# [1] 0.8863081
```

```
qf(0.025, m-1, n-1)
```

```
# [1] 0.2314964
```

```
qf(0.975, m-1, n-1)
```

```
# [1] 5.523407
```

- ▶ The data is consistent with the null hypothesis. There is no evidence to indicate a difference in the variability of heights for boys and girls.

Note that, this test strongly depends on the assumption of normality.



MACQUARIE
University
SYDNEY • AUSTRALIA

Sample Size and Power

$\mu_0 = 19.25$

- ▶ Return to the z-test example in Week 1.
- ▶ Test if a new video encoder has a mean encoder time quicker than the old encoder,

$$H_0 : \mu = 19.25 \quad H_1 : \mu < 19.25$$

- ▶ Summary statistics are $\bar{y} = 18.22$, $n = 10$
 - ▶ Assume $\sigma = 1.5$

$$\begin{aligned} z_{obs} &= \frac{\bar{y} - \mu_0}{\sigma / \sqrt{n}} \\ &= \frac{18.22 - 19.25}{1.5 / \sqrt{10}} \\ &= \frac{-1.03}{0.474} = -2.17 \end{aligned}$$

- ▶ P-Value = $P(Z < -2.17) = 0.015$
- ▶ Evidence that the true mean encoding time for the new encoder is significantly less than 19.25 minutes.

What would have happened if we had obtained the same sample mean of 18.22 but from a sample with a different sample size?

- ▶ Suppose $n = 5$ but $\bar{y} = 18.22$

$$\begin{aligned} z_{obs} &= \frac{\bar{y} - \mu_0}{\sigma/\sqrt{n}} \\ &= \frac{18.22 - 19.25}{1.5/\sqrt{5}} \\ &= \frac{-1.03}{0.67} = -1.535 \end{aligned}$$

- ▶ P-Value = $P(Z < -1.535) = 0.062$.
- ▶ Not enough evidence to reject at the 5% level.
 - ▶ Can't conclude that there is evidence that the new encoder has a quicker mean encoding time.

- ▶ Suppose $n = 1,000$

$$\begin{aligned}Z_{obs} &= \frac{\bar{y} - \mu_0}{\sigma/\sqrt{n}} \\&= \frac{18.22 - 19.25}{1.5/\sqrt{1000}} \\&= \frac{-1.03}{0.0474} = -21.73\end{aligned}$$

- ▶ P-Value = $P(Z < -21.73) = 3.279278 \times 10^{-98} \approx 0$
 - ▶ R loses accuracy at 10^{-15}
- ▶ Reject H_0 at any level of significance (very small P-Value).
- ▶ We are very confident (with a very high confidence level) that the new encoder has a decreased encoding time

- ▶ The larger the sample size, the more confident we can be about the sample mean as an estimate of μ .
- ▶ For very small samples, it is often difficult to obtain a significant result even for large observed differences.
- ▶ For very large samples, you can very often obtain a significant result even for very small (sometimes trivial) observed differences.
- ▶ We can work out, **in advance**, what difference (or \bar{y}) will be significant, given a particular sample size.
- ▶ For a one-tailed test (eg, shorter encoding time) at 5% sig. level, we will reject H_0 if z obs is less than -1.645.

Finding critical value at $n = 1000$

- ▶ Given that

$$z_{obs} = \frac{\bar{y} - \mu_0}{\sigma/\sqrt{n}} = \frac{\bar{y} - 19.25}{1.5/\sqrt{1000}}$$

will be significant if $z_{obs} < -1.645$

- ▶ The minimum value of \bar{y} that achieves this is,

$$\bar{y} < 19.25 - 1.645 \times \frac{1.5}{\sqrt{1000}} = 19.17$$

- ▶ Any observed sample mean less than 19.17 or 0.08 units lower than 19.25 will be deemed significantly less than 19.25 at the 5% level of significance.

Practical Significance vs Statistical Significance



Does a difference of 0.08 units mean anything?

- ▶ If a result is *statistically significant*, we are saying that we are reasonably confident that the true mean is different from the hypothesized value.
 - ▶ However the **magnitude** of that difference may or may not be of interest to us.
- ▶ In this case, the units were minutes.
 - ▶ 0.08 minutes is 4.8 seconds.
- ▶ Is this 4.8 seconds of *practical* use?
- ▶ Before conducting an experiment, the *meaningful* difference should be decided.
 - ▶ This is the *practical significance* amount.
 - ▶ Depends on context.
- ▶ Denote δ as the minimum *practical significance* that would be used.

What difference in video encoding time is considered worthy or important in reality?

- ▶ Suppose marketing decide a new video encoder can only be competitive if the average decrease in encoding time is 0.5 minutes (30 seconds) faster.
 - ▶ They require a $\delta = 0.5$
- ▶ If we observe a difference of less than 0.5 minutes, we don't care if its significant or not.
 - ▶ Customers won't care if the encoding time is less than a minute quicker, so why bother?
 - ▶ A difference that small is negligible.
- ▶ We aim to marry the two concepts. Find a required *sample size* such that we can detect a **pratically significant** result at the 5% level of **statistical significance**.

- ▶ Suppose a \bar{y} of at least 0.5 minutes less than the hypothesized mean (19.25) is observed, $\bar{y} < 18.75$,

What does sample size n have to be, to be able to declare such a result as significant?

- ▶ We want the minimal difference $\delta = 0.5$ to satisfy

$$z_{obs} = \frac{\bar{y} - \mu}{\frac{1.5}{\sqrt{n}}} < -1.645$$

- ▶ Rearranging we have,

- ▶ $\frac{-1.645 \times 1.5}{\sqrt{n}} > -0.5 (= 18.75 - 19.25) \Rightarrow \sqrt{n} > \frac{-1.645 \times 1.5}{-0.5} = 4.935$

- ▶ $n > (4.935)^2 = 24.35423$.

- ▶ Need at least 25 observations in our sample data set to be able to detect an encoding time at least 30 seconds quicker than 19.25 minutes and be able to conclude at the 5% level of significance.

- ▶ If we have a sample size of less than 25 and we observe a decrease of less than 0.5 minutes
 - ▶ the result won't be *statistically* significant at the 5% level.

- ▶ **General result** for sample size calculations.

$$n > \left(\frac{z_{critical} \times \sigma}{\delta} \right)^2$$

- ▶ $z_{critical}$ takes the values
 - ▶ z_{α} for a one-sided test.
 - ▶ $z_{\alpha/2}$ for a two-sided test.
 - ▶ δ is the minimal difference required.
- ▶ **Note:** σ can be replaced with s and consequently z_{crit} replaced with the appropriate $t_{\nu, crit}$.
 - ▶ More computation of sample sizes can be viewed in STAT3195.

For any test:

- ▶ Type I error = Rejecting H_0 when H_0 is true.
 - ▶ $P(\text{Type I error}) = \alpha$, the significance level.
- ▶ Type II error = Not rejecting H_0 when H_0 is false.
 - ▶ $P(\text{Type II error}) = \beta$, which depends on:
 - ▶ α .
 - ▶ The true value of the parameter.
 - ▶ Whether the test is one or two-sided.
- ▶ Power of the test = $P(\text{Rejecting } H_0 \text{ when } H_0 \text{ is false}) = 1 - \beta$.
- ▶ We can calculate the *Power* of a test for any given set of circumstances.

- ▶ This time we want test the following new hypotheses:

$$H_0 : \mu = 17.25 \quad (\text{different value})$$

$$H_1 : \mu > 17.25$$

- ▶ Suppose $\alpha = 0.05$ and $\sigma = 1.5$.
- ▶ Use the same summary statistics: $\bar{y} = 18.22$ and $n = 10$.
- ▶ As Power of the test is related to us rejecting H_0 , we have to first figure out when the rejection going to occur based on \bar{y} .
- ▶ Can compute the critical value/the minimum \bar{y} that will give a statistically significant result at the 5% level by

$$\frac{\bar{y} - 17.25}{\frac{1.5}{\sqrt{10}}} > 1.645 = z_{1-\alpha} = z_{0.95}$$

- ▶ Solving gives $\bar{y} > 18.03$
 - ▶ For any sample of size 10, H_0 will be rejected if we get $\bar{y} > 18.03$.

- ▶ If we carry out a one-sided z-test at the 5% level of significance to test

$$H_0 : \mu = 17.25 \quad H_1 : \mu > 17.25$$

with a sample size of 10.

- ▶ For our statistical procedure, we will
 - ▶ Reject H_0 if $\bar{y} > 18.03$
 - ▶ Not reject H_0 if $\bar{y} < 18.03$
- ▶ We can work out the probability of getting $\bar{y} > 18.03$ for particular values of μ
 - ▶ This is the **Power** of test for various values of μ .
- ▶ This is carried out at the design/plan stage of a study (eg, an experiment). At this time, we may have no idea
 - ▶ what the true value of μ is or
 - ▶ what value of \bar{y} will be observed.

- ▶ Suppose the true mean is $\mu = 18.5$

Power of the test when " $\mu = 18.5$ "

$$= P(\bar{Y} > 18.03, \text{ given } \mu = 18.5)$$

$$= P\left(\frac{\bar{Y} - 18.5}{1.5/\sqrt{10}} > \frac{18.03 - 18.5}{1.5/\sqrt{10}}\right)$$

$$= P(Z > -0.99)$$

$$= 0.84$$

- ▶ If the **true mean** is 18.5, there is an 84% chance that $H_0 : \mu = 17.25$ is rejected in favour of $H_1 : \mu > 17.25$!
 - ▶ On the other hand, there is a 16% chance H_0 would not be rejected.
 - ▶ Making a Type II error.

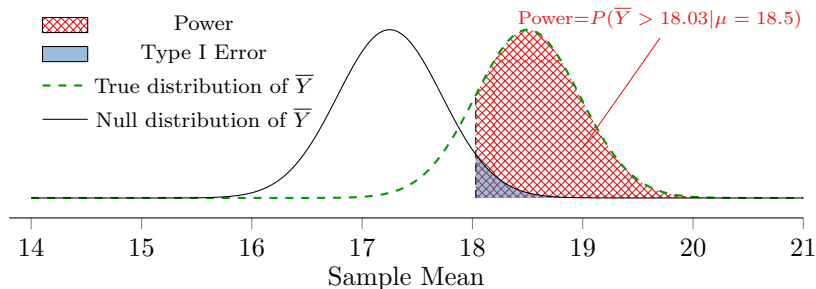


Figure 3: Plot of the distribution of the sample mean when $\mu = 18.5$ or $H_0 : \mu = 17.25$

Power at $\mu = 18.0$

$$\begin{aligned} &= P(\bar{Y} > 18.03 | \mu = 18.0) \\ &= P\left(Z > \frac{18.03 - 18.0}{1.5/\sqrt{10}}\right) \\ &= P(Z > 0.063) \\ &= 0.476 \end{aligned}$$

Power at $\mu = 17.25$

$$\begin{aligned} &= P(\bar{Y} > 18.03 | \mu = 17.25) \\ &= P\left(Z > \frac{18.03 - 17.25}{1.5/\sqrt{10}}\right) \\ &= P(Z > 1.645) \\ &= 0.05, \text{ as expected} \end{aligned}$$

- ▶ If the **true mean** is 18.0, there is an 48% chance that $H_0 : \mu = 17.25$ is rejected in favour of $H_1 : \mu > 17.25$!
- ▶ If the **true mean** is 17.25, there is an 5% chance that $H_0 : \mu = 17.25$ is rejected in favour of $H_1 : \mu > 17.25$!

Power vs Possible values of μ

- ▶ For testing $H_0 : \mu = 17.25$ against $H_1 : \mu > 17.25$,
- ▶ $n = 10$ and $\alpha = 0.05$, 5% significance

```
mu0 = 17.25
mu = c(16.25, 17.25, 18, 18.5, 19)
mbound = qnorm(0.95) * 1.5/sqrt(10) + 17.25
power = 1 - pnorm((mbound - mu)/(1.5/sqrt(10)))
powertab = cbind(mu, mu - 17.25, power)
colnames(powertab) = c("$\\mu$", "$\\mu - 17.25$", "Power")
kable(powertab)
```

μ	$\mu - 17.25$	Power
16.25	-1.00	0.0000874
17.25	0.00	0.0500000
18.00	0.75	0.4745987
18.50	1.25	0.8390052
19.00	1.75	0.9795464

- ▶ **Note:** The greater the distance between the *true* mean and the mean claimed in H_0 , the higher the power is to detect that difference.

- ▶ If we carry out a two-tailed test, the same principles apply but we have *two rejection regions*.
- ▶ Reject H_0 at the 5% level if $|z_{obs}| > 1.96$.
 - ▶ Either $z_{obs} < -1.96$ or $z_{obs} > 1.96$.
 - ▶ Either $\bar{y} < 17.25 - 1.96 \times 1.5/\sqrt{10} = 16.23$.
 - ▶ Or $\bar{y} > 17.25 + 1.96 \times 1.5/\sqrt{10} = 18.18$.
- ▶ Power = $P(\bar{Y} > 18.18) + P(\bar{Y} < 16.32)$ for the chosen true μ .

- ▶ Q1: 20 participants are randomly selected by their GP to assess the effects of eating fruit on blood sugar levels (mmol/L). Their blood sugar levels were measured for each participant upon entering the trial and again one hour later after eating three pieces of fruit.
 - ▶ A: Z test
 - ▶ B: One sample t test
 - ▶ C: Paired t test
 - ▶ D: Independent two sample t test
 - ▶ E: Modified (Welch) two sample t test

- ▶ Q2: A study was conducted to determine whether the mean birth weight of Aboriginal babies is lower than that of non-Aboriginal babies. The birth weights of all Aboriginal and non-Aboriginal babies born in NSW hospitals were recorded and analysed. Assume the variances are equal and the true population variance is unknown.
- ▶ A: Z test
 - ▶ B: One sample t test
 - ▶ C: Paired t test
 - ▶ D: Independent two sample t test
 - ▶ E: Modified (Welch) two sample t test

- ▶ Q3: A study was conducted to determine if the average price of coffee in Sydney CBD is higher or lower than \$3.50 (the average price of coffee in Melbourne CBD, say). A total of 1000 adults, male and female, were randomly selected and the cost of their morning coffee recorded. Assume the variances are equal and the true population variance is unknown.
 - ▶ A: Z test
 - ▶ B: One sample t test
 - ▶ C: Paired t test
 - ▶ D: Independent two sample t test
 - ▶ E: Modified (Welch) two sample t test