Biostatistics

Chapter 5B Hypothesis Testing : One or Two samples

Jing Li

jing.li@sjtu.edu.cn

http://cbb.sjtu.edu.cn/~jingli/courses/2017fall/bi372/

Dept of Bioinformatics & Biostatistics, SJTU

Review Questions (5 min)

- How to increase statistics power ?
- What's 99 % confidence interval for a mean?

Topics

- Z-test
  - Testing single population mean
  - Testing sample proportions
- One sample T-test
- Comparison of Two means t-test
  - Paired t-test
  - Independent samples

Topics

- Z-test
  - Testing single population mean
  - Testing sample proportions
- One sample T-test
- Comparison of Two means t-test
  - Paired t-test
  - Independent samples

Recall

- When the sample size is small (approximately < 100) then the Student's *t* distribution should be used.
- The test statistic is known as "*t*".
- The curve of the *t* distribution is flatter than that of the Z distribution but as the sample size increases, the *t*-curve starts to resemble the Z-curve.

Degrees of Freedom

- The curve of the t distribution varies with sample size (the smaller the size, the flatter the curve)
- In using the t-table, we use "degrees of freedom" based on the sample size.
- For a one-sample test, df = n – 1.
- When looking at the table, find the t-value for the appropriate df = n-1. This will be the cutoff point for your critical region.

Formula for one sample t-test:

$$t = \frac{\overline{X} - \mu}{S / \sqrt{n}}$$

The normality assumption…

- T-tests (and all linear models, in fact) have a "normality assumption":
  - If the outcome variable is not normally distributed and the sample size is small, a t-test is inappropriate

The normality assumption…

- If the underlying data are not normally distributed AND n is small**, the means do not follow a t-distribution (so using a t-test will result in erroneous inferences).
- Data transformation or non-parametric tests should be used instead.
- **How small is too small? No hard and fast rule—depends on the true shape of the underlying distribution.

Single population mean (large n)

- Hypothesis test:

$$Z = \frac{\text{observed mean} - \text{null mean}}{\frac{s}{\sqrt{n}}}$$

- Confidence Interval

$$\text{confidence interval} = \text{observed mean} \pm Z_{\alpha/2} * (\frac{s}{\sqrt{n}})$$

Single population mean (small n, normally distributed)

- Hypothesis test:

$$T_{n-1} = \frac{\text{observed mean} - \text{null mean}}{\frac{s}{\sqrt{n}}}$$

- Confidence Interval

$$\text{confidence interval} = \text{observed mean} \pm \mathrm{T}_{n-1,\alpha/2} * (\frac{s}{\sqrt{n}})$$

Practice

- In the population, the average IQ is 100 with a standard deviation of 15. A team of scientists wants to test a new medication to see if it has either a positive or negative effect on intelligence, or no effect at all. A sample of 30 participants who have taken the medication has a mean of 140. Did the medication affect intelligence, using alpha = 0.05?

Comparison of Two Group

- Are the population means different? (continuous data)
- Paired design
  - Before-after data
  - Twin data
  - Matched case-control
- Two independent sample design
  - Randomized trial
  - Smokers to non-smokers

Paired Design—Example: Before vs. After

- A study I analyzed was a tumor size study. Having an accurate measure of tumor size is extremely important because it allows a physician to accurately determine if a tumor is growing, shrinking or remaining constant.
- The problem is that often the measurements of the tumor size vary from physician to physician.
- In the past, tumor size was measured using the linear distance across the tumor, but this was found to be very variable because of the irregular shape of some tumors. A new method called the RECIST criteria, which traces the outside of the tumor, measures the volume of the tumor. The volumetric method was believed to give more consistent measures of the volume of the tumor.

Available data

- For a portion of the study, a pair of doctors were shown the same set of tumor pictures. The volume of the tumor was measured by two separate physicians under similar conditions.
- Question of interest: did the measurements from the two physicians significantly differ?
- If not, then there would be no evidence that the volume measurements change based on physician.

Data

- 20 scans were measured by each physician (10 are shown here)
- Measurements in cm^3
- What can you say about these samples?
  - Two measurement on the same person
  - They are related so we must account for this

Tumor	Dr. 1	Dr. 2
1	15.8	17.2
2	22.3	20.3
3	14.5	14.2
4	15.7	18.5
5	26.8	28.0
6	24.0	24.8
7	21.8	20.3
8	23.0	25.4
9	29.3	27.5
10	20.5	19.7

Difference of Two Groups

- We can measure the effect of the treatment in each person by taking the difference

$$d_i = x_{1i} - x_{2i}$$

- Instead of having two samples, we can consider our dataset to be one sample of differences
  - **Just like the one sample problem**

Tumor	Dr. 1	Dr. 2	Difference
1	15.8	17.2	-1.4
2	22.3	20.3	2.0
3	14.5	14.2	0.3
4	15.7	18.5	-2.8
5	26.8	28.0	-1.2
6	24.0	24.8	-0.8
7	21.8	20.3	1.5
8	23.0	25.4	-2.4
9	29.3	27.5	1.8
10	20.5	19.7	0.8

Difference of Two Groups

- Volume from Dr. 1
  - Population mean: μ_1
  - Sample mean: $\bar{x}_1$
- Volume from Dr. 2
  - Population mean: μ_2
  - Sample mean: $\bar{x}_2$
- Difference
  - Population mean: $\delta = \mu_1 - \mu_2$
  - Sample mean: $\bar{d} = \frac{\sum_{i=1}^{n} d_i}{n}$

Difference of Two Groups

- Assuming the differences are normally distributed, can use t-distribution with n-1 df where n is the number of differences

$$t = \frac{\bar{d} - \delta}{s_d / \sqrt{n}}$$

- Standard deviation of differences

$$s_d = \sqrt{\frac{\sum_{i=1}^{n} \left(d_i - \bar{d}\right)^2}{n-1}}$$

- Test statistic acts just like one sample （T-table）

Paired t-test

1) Two dependent samples; alpha=0.05
2) Null hypothesis: No difference between physicians effect

$$H_0 : \mu_{dr1} = \mu_{dr2} \Rightarrow \delta = \mu_{dr1} - \mu_{dr2} = 0$$

3) Test statistic: t-statistic with df $t = \frac{\bar{d} - \delta}{s_d / \sqrt{n}} = \frac{-0.24}{1.66/\sqrt{20}} = -0.646$
4) p-value=0.53
5) Fail to reject null hypothesis
6) Conclusion: there is no evidence of a difference in tumor volume measurement based on physician

T-table

PERCENTAGE POINTS OF THE T DISTRIBUTION

		Tail Probabilities							
		One Tail	0.10	0.05	0.025	0.01	0.005	0.001	0.0005	
		Two Tails	0.20	0.10	0.05	0.02	0.01	0.002	0.001	
D	1		3.078	6.314	12.71	31.82	63.66	318.3	637	1
E	2		1.886	2.920	4.303	6.965	9.925	22.330	31.6	2
G	3		1.638	2.353	3.182	4.541	5.841	10.210	12.92	3
R	4		1.533	2.132	2.776	3.747	4.604	7.173	8.610	4
E	5		1.476	2.015	2.571	3.365	4.032	5.893	6.869	5
E	6		1.440	1.943	2.447	3.143	3.707	5.208	5.959	6
S	7		1.415	1.895	2.365	2.998	3.499	4.785	5.408	7
	8		1.397	1.860	2.306	2.896	3.355	4.501	5.041	8
O	9		1.383	1.833	2.262	2.821	3.250	4.297	4.781	9
F	10		1.372	1.812	2.228	2.764	3.169	4.144	4.587	10
	11		1.363	1.796	2.201	2.718	3.106	4.025	4.437	11
F	12		1.356	1.782	2.179	2.681	3.055	3.930	4.318	12
R	13		1.350	1.771	2.160	2.650	3.012	3.852	4.221	13
E	14		1.345	1.761	2.145	2.624	2.977	3.787	4.140	14
E	15		1.341	1.753	2.131	2.602	2.947	3.733	4.073	15
D	16		1.337	1.746	2.120	2.583	2.921	3.686	4.015	16
O	17		1.333	1.740	2.110	2.567	2.898	3.646	3.965	17
M	18		1.330	1.734	2.101	2.552	2.878	3.610	3.922	18
	19		1.328	1.729	2.093	2.539	2.861	3.579	3.883	19
	20		1.325	1.725	2.086	2.528	2.845	3.552	3.850	20
	21		1.323	1.721	2.080	2.518	2.831	3.527	3.819	21
	22		1.321	1.717	2.074	2.508	2.819	3.505	3.792	22
	23		1.319	1.714	2.069	2.500	2.807	3.485	3.768	23
	24		1.318	1.711	2.064	2.492	2.797	3.467	3.745	24
	25		1.316	1.708	2.060	2.485	2.787	3.450	3.725	25
	26		1.315	1.706	2.056	2.479	2.779	3.435	3.707	26
	27		1.314	1.703	2.052	2.473	2.771	3.421	3.690	27
	28		1.313	1.701	2.048	2.467	2.763	3.408	3.674	28
	29		1.311	1.699	2.045	2.462	2.756	3.396	3.659	29
	30		1.310	1.697	2.042	2.457	2.750	3.385	3.646	30
	32		1.309	1.694	2.037	2.449	2.738	3.365	3.622	32
	34		1.307	1.691	2.032	2.441	2.728	3.348	3.601	34

Paired t-test

1) Two dependent samples; alpha=0.05
2) Null hypothesis: No difference between physicians effect

$$H_0 : \mu_{dr1} = \mu_{dr2} \Rightarrow \delta = \mu_{dr1} - \mu_{dr2} = 0$$

3) Test statistic: t-statistic with df $t = \frac{\bar{d} - \delta}{s_d / \sqrt{n}} = \frac{-0.24}{1.66/\sqrt{20}} = -0.646$

4) p-value>0.05, fail to reject null hypothesis
5) Conclusion: there is no evidence of a difference in tumor volume measurement based on physician

Another example

- Ten non-pregnant women 16–49 years old who were beginning a regimen of oral contraceptive (OC) use had their blood pressures measured prior to starting OC use and three-months after consistent OC use.
- The goal of this small study was to see what, if any, changes in average blood pressure were associated with OC use in such women
- The data on the following slides shows the resulting pre- and post-OC use systolic BP measurements for the 10 women in the study

Data

	BP Before OC	BP After OC	After-Before
1.	115	128	13
2.	112	115	3
3.	107	106	-1
4.	119	128	9
5.	115	122	7
6.	138	145	7
7.	126	132	6
8.	105	109	4
9.	104	102	-2
10.	115	117	2

$\bar{x}_{bef\ ore} = 115.6$ $\bar{x}_{af\ ter} = 120.4$ $\bar{x}_{dif\ f} = 4.8$

Hypothesis test

- **Null:** typically represents the hypothesis that there is "no association" or "no difference"
  - For example, there is no association between oral contraceptive use and blood pressure
    - H_o: $\mu = 0$
- **Alternative:** the very general complement to the null
  - For example, there is an association between blood pressure and oral contraceptive use
    - H_A: $\mu \neq 0$

Hypothesis test

- We are testing both hypotheses at the same time
  - Our result will allow us to either:
    - "Reject H_0"
      or
    - "Fail to reject H_o"

- We start by assuming the null (H_o) is true, and asking:
  - How likely is the result we got from our sample if H_o is the truth —i.e., no change in mean blood pressure after taking OCs?
  - $\bar{x}$ would have to be far from zero to claim H_A is true
    - But is $\bar{x}$ = 4.8 mmHg big enough to choose H_A?

Sampling distribution

Sampling distribution of the sample mean is the (theoretical) distribution of all possible values of $\bar{x}$ from samples of same size,

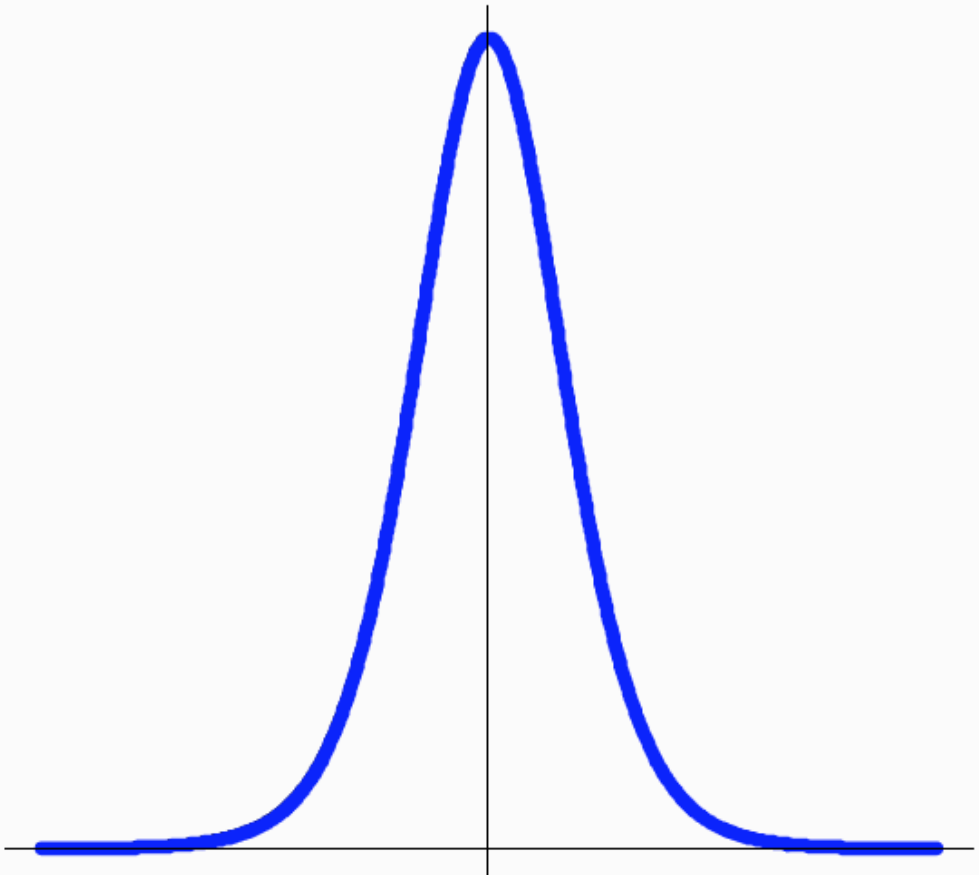
For BP example, theory tells us it is a t_9 *distribution*

Recall, the sampling distribution is centered at the "truth," the underlying value of the population mean, μ

- In hypothesis testing, we start under the assumption that H_0 is true—so the sampling distribution under this assumption will be centered at μ_0, the null mean

Sampling distribution

- Sampling distribution of sample mean differences (after-before) in BP, from samples of size n=10



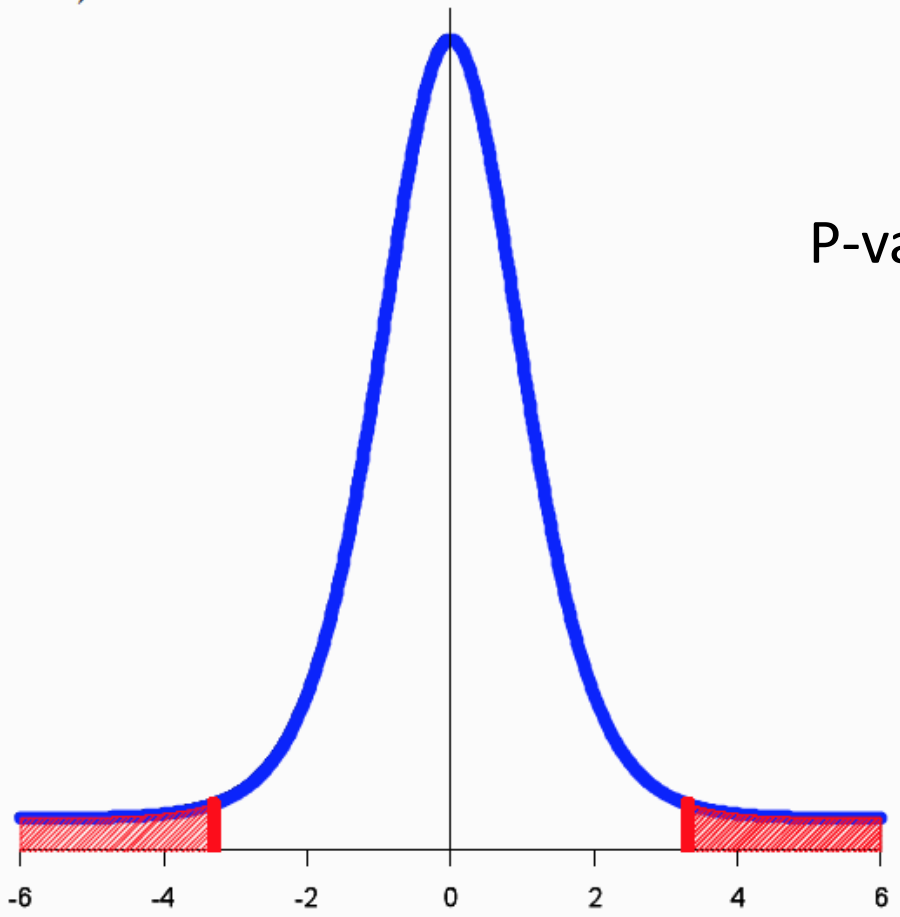
t-test

$$t = \frac{\bar{x}_{diff} - 0}{S\hat{E}(\bar{x})}$$

$$t = \frac{4.8 - 0}{4.6/\sqrt{10}} = \frac{4.8}{1.45} \approx 3.3$$

t-test

- The p-value is the probability of getting a sample result as (or more) extreme than what you observed (3.3) away from $\mu_0 = 0$ (in either direction from 0)

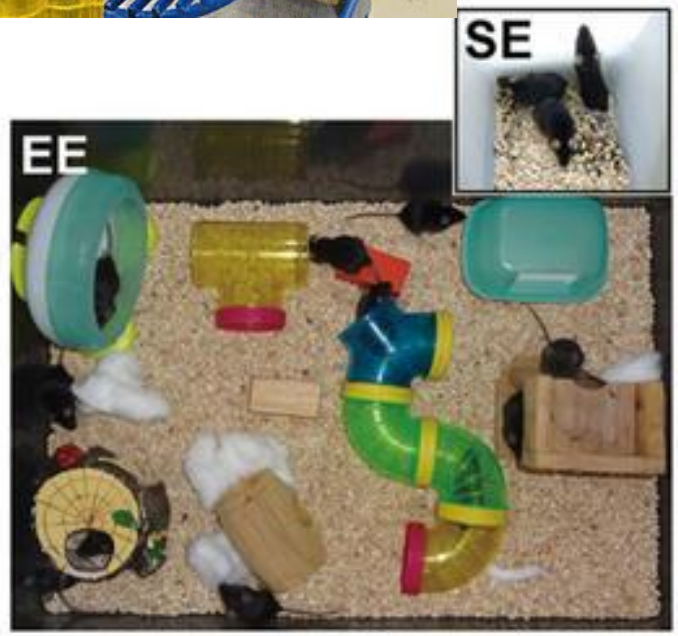


P-value <0.05

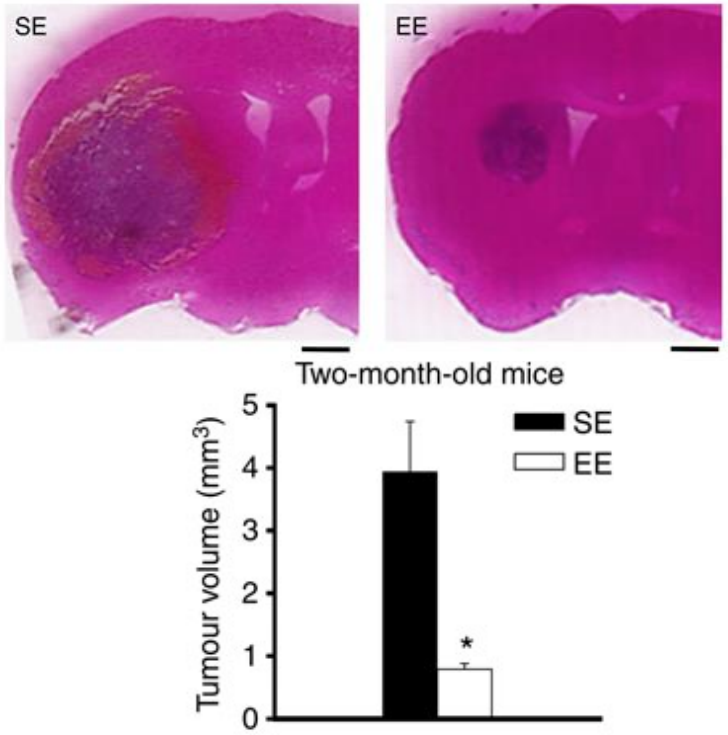
Extensions

- Some additional examples of paired samples are:
  - **Differences between left and right eye**
  - **Matched samples**
  - **Other example ?**

Example (Published 30 March 2015)



Enriched environment reduces glioma growth



Effect of SE or EE housing on 2-month-old mice transplanted with glioma (*n*=5 per group; ***P*<0.05, Student's *t*-test**)

*Nature Communications, 2015, **6**, 6623*

Comparison of Two Group

- Paired design
  - Before-after data
  - Twin data
  - Matched case-control
- Two independent sample design
  - Randomized trial
  - Smokers to non-smokers

Unpaired samples

- Often it is impractical to design study to use the same patients for both group
  - Comparison of cholesterol in males and females
  - Time constraints
- Since the samples are not paired, we cannot use the difference between the individual samples

Example

- Another aspect of the tumor volume study was trying to compare the tumor volume among patients with different cancer. The average tumor size is important to know the effect of treatment can be determined.
- In this study, patients with brain, breast and liver tumors, but initially we will only compare the brain and breast cancers.
- All of the tumors were measured using the RECIST method

Null hypothesis

- The null hypothesis is that there is no difference between the volume of the tumor in the two forms of cancer

$H_0: m_{brain} = m_{breast}$, or $m_{brain} - m_{breast} = 0$

- More generally, we can test if the difference between two groups is a specific value, $m_1 - m_2 = D$
  - This occurs when comparing two treatment groups and we are interested if the two groups are different

Difference in the sample means

- We are going to use the difference of the means as our test statistic, but we need to estimate the variance of this difference to determine if the difference is significant
- Basic form of test statistic:
  - **Standard deviations known** **unknown& small sample size**

$$z = \frac{\left(\overline{x_1} - \overline{x_2}\right) - \left(\mu_1 - \mu_2\right)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \qquad t = \frac{\left(\overline{x_1} - \overline{x_2}\right) - \left(\mu_1 - \mu_2\right)}{s_{\overline{x_1} - \overline{x_2}}}$$

- The estimate of the standard deviation changes when
  - **The samples have equal variance OR**
  - **The samples have unequal variance**

Equal variance

- Sometimes we will be willing to assume that the variance in the two groups is equal:

$$\sigma_1^2 = \sigma_2^2 = \sigma^2$$

- If we know this variance, we can use the z-statistic

- Often we have to estimate s^2 with the sample variance from each of the samples, s_1^2, s_2^2
- Since we have two estimates of one quantity we pool the two estimates

Equal variance continued

- The estimate of s is given by:

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

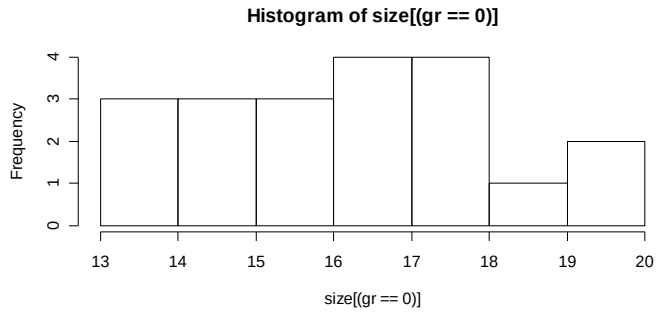
- The t-statistic based on the pooled variance is very similar to the z-statistic as always:

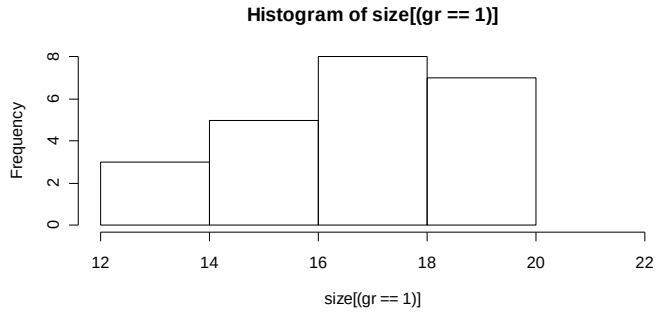
$$t = \frac{(\overline{x_1} - \overline{x_2}) - (\mu_1 - \mu_2)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

- The t-statistic has a t-distribution with degrees of freedom $n_1 + n_2 - 2$

- For the tumor volume study, there were 20 brain cancer subjects and 28 breast cancer subjects
- The summary statistics and histogram for the data are given here
- What can you say about the distributions?
- Does the equal variance assumption seem valid in this case?

	Brain	Breast
n	20	28
xbar	16.2 cm^3	17.5 cm^3
s^2	3.49	6.0





Hypothesis test

1) Two independent samples with equal variance; alpha = 0.05
2) H_0: mean brain tumor size = mean breast tumor size
3) $$t = \frac{(\overline{x_1} - \overline{x_2}) - (\mu_1 - \mu_2)}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{16.2 - 17.5 - 0}{2.23\sqrt{\frac{1}{20} + \frac{1}{28}}} = -2.054$$
4) p-value: 0.046
5) Reject null hypothesis
6) Conclusion: There is a significant difference in the size of brain and breast cancer tumors

Unequal variance

- Often, we are unwilling to assume that the variances are equal
- We now write the test statistic as:

$$t = \frac{\left(\overline{x_1} - \overline{x_2}\right) - \left(\mu_1 - \mu_2\right)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

- The distribution of this statistic is difficult to derive and we approximate the distribution using a t-distribution with ν degrees of freedom

$$\nu = \frac{\left[\left(s_1^2/n_1\right) + \left(s_2^2/n_2\right)\right]^2}{\left[\frac{\left(s_1^2/n_1\right)^2}{\left(n_1 - 1\right)} + \frac{\left(s_2^2/n_2\right)^2}{\left(n_2 - 1\right)}\right]}$$

- This is called the Satterthwaite or Welch approximation
  - When you complete a two-sample t-test in R and the variances are not assumed equal, this approximation is used

Example 2

- For the comparison of the brain cancers to the liver cancers, the variances are much more different.
- Let's use the unequal variance two sample t-test in this case

	Brain	Liver
n	20	17
xbar	16.2 cm^3	19.35 cm^3
s^2	3.49	14.4

Example 2

1) Two independent samples with unequal variance; alpha = 0.05

2) H_0: mean brain tumor size = mean liver tumor size

3) $$t = \frac{(\overline{x_1} - \overline{x_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{16.15 - 19.35 - 0}{\sqrt{\frac{3.5}{20} + \frac{14.4}{17}}} = -3.17$$

4) p-value: 0.0044

5) Reject null hypothesis

6) Conclusion: There is a significant difference in the size of the brain and liver tumor size

Can we test if the variances are equal?

- Since we can never be sure if the variances are equal, could we test if they are equal?
- Of course we can!!!
  - But, remember there is error in every statistical test
  - Sometimes it is just preferred to use the unequal variance unless there is a good reason

Equality of variance

- $H_0: s_1^2 = s_2^2$
- To test this hypothesis, we use the sample variances:

  s_1^2, s_2^2

- If one of the variances is much larger than the other, this is evidence against the null

Test of equality

- One way to test if the two variances are equal is to check if the ratio is equal to 1
- Under the null, the ratio simplifies to $\frac{s_1^2}{s_2^2}$

- The ratio of 2 chi-square random variables has an F-distribution
- The F-distribution is defined by the numerator and denominator degrees of freedom
- Here we have an F-distribution with n_1-1 and n_2-1 degrees of freedom
- This works better with $s_1^2 > s_2^2$

Sample size for paired data:

$$n = \frac{\sigma_d^{\,2}(Z_\beta + Z_{\alpha/2})^2}{\text{difference}^2}$$

where:

n = sample size

σ = standard deviation of the within-pair difference

difference = clinically meaningful difference

Z_β = corresponds to power (.84 = 80% power)

$Z_{\alpha/2}$ = corresponds to two-tailed significance level (1.96 for $\alpha = .05$)

Sample size for Unpaired data:

$$n_1 = \frac{(r+1)}{r} \frac{\sigma^2 (Z_\beta + Z_{\alpha/2})^2}{\text{difference}^2}$$

where:

n_1 = size of smaller group

r = ratio of larger group to smaller group

σ = standard deviation of the characteristic

difference = clinically meaningful difference in means of the outcome

Z_β = corresponds to power (.84 = 80% power)

$Z_{\alpha/2}$ = corresponds to two-tailed significance level (1.96 for $\alpha = .05$)

Practice

In fall 2004, students in the 2 p.m. section of my Biological Data Analysis class had an average height of 66.6 inches, while the average height in the 5 p.m. section was 64.6 inches. Are the average heights of the two sections significantly different? Here are the data:

2 p.m.	5 p.m.
69	68
70	62
66	67
63	68
68	69
70	67
69	61
67	59
62	62
63	61
76	69
59	66
62	62
62	62
75	61
62	70
72	
63	