Lecture 5

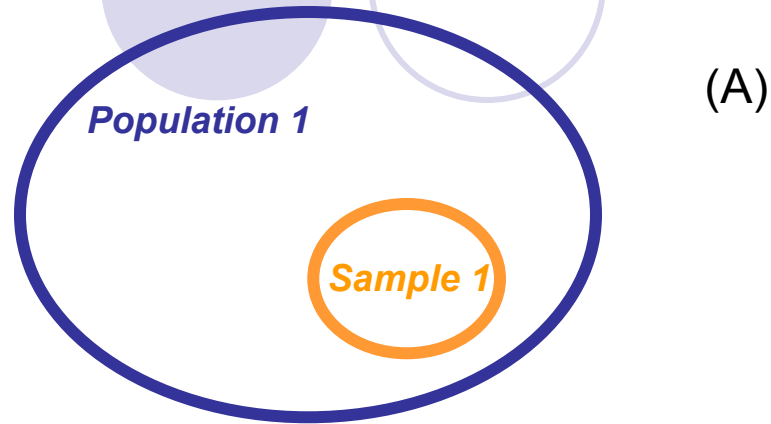## Two population tests of Means and Proportions
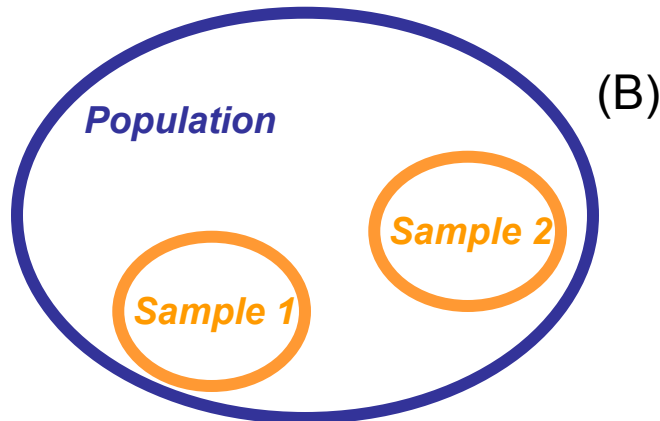
Section 7.2. (objectives)

- **Comparing two means**

  - Two-sample *z* distribution
  - Two independent samples *t*-distribution
  - Two sample *t*-test
  - Two-sample *t*-confidence interval
  - Robustness
  - Details of the two sample *t* procedures

Comparing two samples

Population 1

Sample 1

(A)

Population 2

Sample 2

Which

is it?

Population

(B)

Sample 2

Sample 1

Independent samples: Subjects in one samples are completely unrelated to subjects in the other sample.

We often compare two treatments used on **independent** samples.

Is the difference between both treatments due only to variations from the random sampling (B), or does it reflect a true difference in population means (A)?

Two-sample z distribution (std dev is assumed to be known)

We have **two independent SRSs** (simple random samples) coming maybe from two distinct populations with (μ_1, σ_1) and (μ_2, σ_2). We use $\overline{x}_1$ and $\overline{x}_2$ to estimate the unknown μ_1 and μ_2.

When both populations are normal, the sampling distribution of $(\overline{x}_1 - \overline{x}_2)$ is also normal, with standard deviation :

$$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Then the **two-sample z statistic** has the standard normal *N(0, 1)* sampling distribution.

$$z = \frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Inference – Two Populations, Known Population Standard Deviations

Alternative Hypothesis	P-value
H_a: $\mu_1 > \mu_2$	$P(Z>z)$
H_a: $\mu_1 < \mu_2$	$P(Z<z)$
H_a: $\mu_1 \neq \mu_2$	$2*P(Z>| z |)$

Two independent samples *t* distribution (the population's std. dev. is not known)

We have **two independent SRSs** (simple random samples) coming

maybe from two distinct populations with (μ_1, σ_1) and (μ_2, σ_2) unknown.

We use $(\bar{x}_1, s_1)$ and $(\bar{x}_2, s_2)$ to estimate (μ_1, σ_1) and (μ_2, σ_2) respectively.

To compare the means, both populations should be normally

distributed. However, in practice, it is enough that the two distributions

have similar shapes and that the sample data contain no strong outliers.

The two-sample *t* statistic follows approximately the t distribution with a standard error SE (spread) reflecting variation from both samples:

$$SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Conservatively, the degrees of freedom is equal to the smallest of $(n_1 - 1, n_2 - 1)$.

t distribution, degrees of freedom **df**

$$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$\mu_1 - \mu_2$

$\overline{x}_1 - \overline{x}_2$
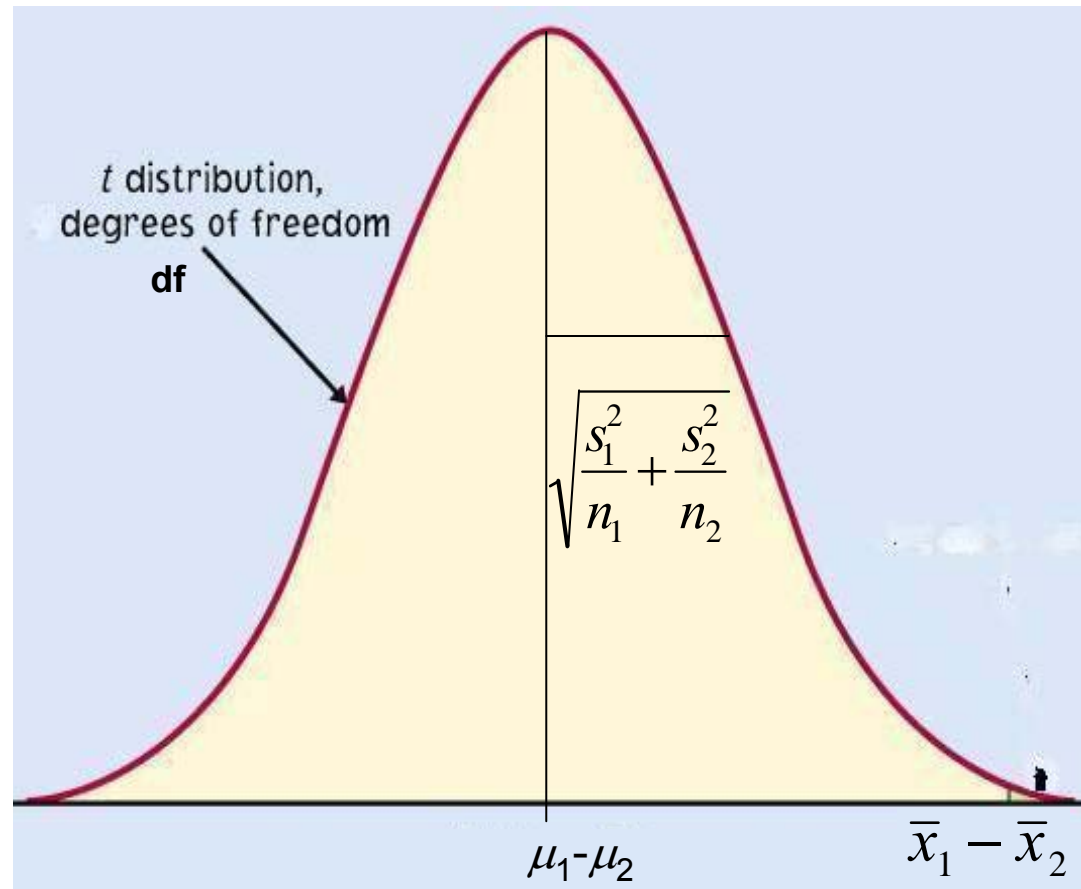
Two-sample *t*-test

The null hypothesis is that both population means μ_1 and μ_2 are equal, thus their difference is equal to zero.

$$H_0: \mu_1 = \mu_2 \iff \mu_1 - \mu_2 = 0$$

with either a one-sided or a two-sided alternative hypothesis.

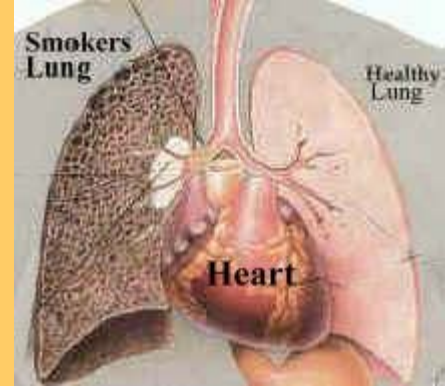We find how many standard errors (SE) away from $(\mu_1 - \mu_2)$ is ($\bar{x}_1 - \bar{x}_2$) by standardizing with *t*:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{SE}$$

Because in a two-sample test H_0 is $(\mu_1 - \mu_2) = 0$, we simply use

With df = smallest($n_1 - 1$, $n_2 - 1$)

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$$

Does smoking damage the lungs of children exposed to parental smoking?

Forced vital capacity (FVC) is the volume (in milliliters) of air that an individual can exhale in 6 seconds.
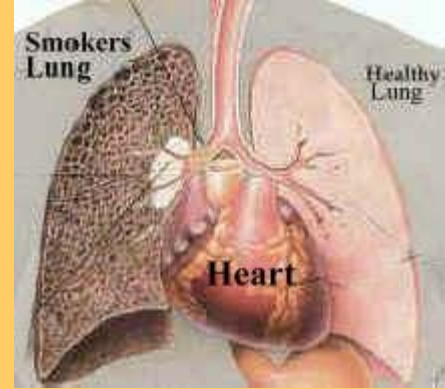
FVC was obtained for a sample of children not exposed to parental smoking and a group of children exposed to parental smoking.

Parental smoking	FVC $\bar{x}$	s	n
Yes	75.5	9.3	30
No	88.2	15.1	30

We want to know whether parental smoking decreases children's lung capacity as measured by the FVC test.

Is the mean FVC lower in the population of children exposed to parental smoking?

H_0: $\mu_{smoke} = \mu_{no}$ <=> $(\mu_{smoke} - \mu_{no}) = 0$

H_a: $\mu_{smoke} < \mu_{no}$ <=> $(\mu_{smoke} - \mu_{no}) < 0$ (one sided)

The difference in sample averages follows approximately the *t* distribution: $t\left(0, \sqrt{\dfrac{s^2_{smoke}}{n_{smoke}} + \dfrac{s^2_{no}}{n_{no}}}\right)$, df 29

We calculate the *t* statistic:

$$t = \frac{\bar{x}_{smoke} - \bar{x}_{no}}{\sqrt{\dfrac{s^2_{smoke}}{n_{smoke}} + \dfrac{s^2_{no}}{n_{no}}}} = \frac{75.5 - 88.2}{\sqrt{\dfrac{9.3^2}{30} + \dfrac{15.1^2}{30}}}$$

Parental smoking	FVC $\bar{x}$	s	n
Yes	75.5	9.3	30
No	88.2	15.1	30

$$t = \frac{-12.7}{\sqrt{2.9 + 7.6}} \approx -3.9$$

In table C, for df 29 we find:

$|t| > 3.659$ => $p < 0.0005$ (one sided)

It's a very significant difference, we reject H_0.

**Lung capacity is significantly impaired in children of smoking parents.**

Two sample *t*-confidence interval

Because we have two independent samples we use the difference between both sample averages $(\overline{x}_1 - \overline{x}_2)$ to estimate $(\mu_1 - \mu_2)$.

Practical use of *t*: *t**

- *C* is the area between $-t^*$ and t^*.

- We find t^* in the line of Table C for df = smallest $(n_1-1; n_2-1)$ and the column for confidence level C.

- The margin of error *m* is:

$$m = t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = t^* \cdot SE$$

$$SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

C

m   m

$-t^*$ t^*
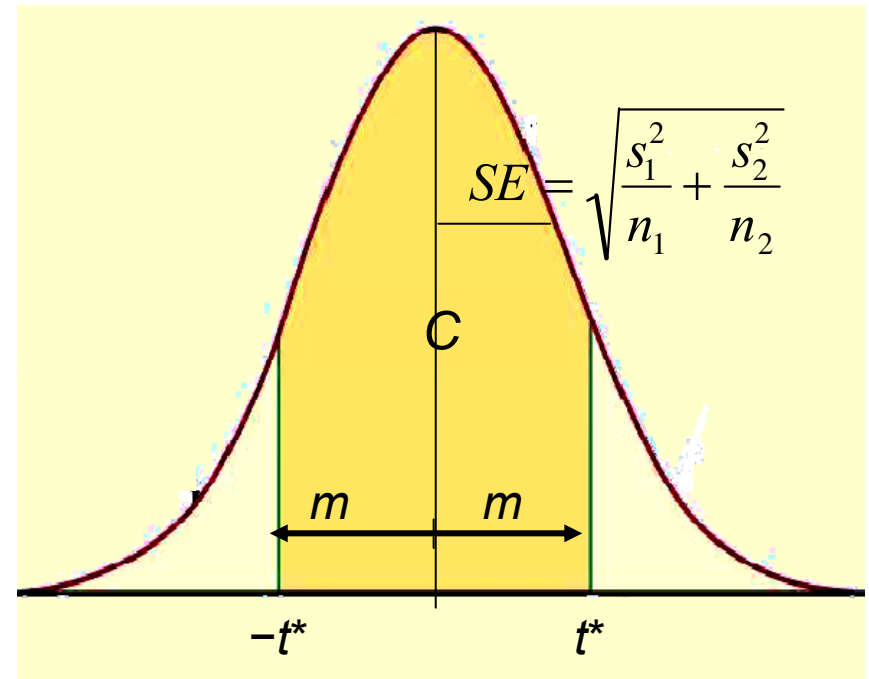
Common mistake !!!

A common mistake is to calculate a one-sample confidence interval for μ_1 and then check whether μ_2 falls within that confidence interval, or vice-versa.

This is WRONG because the variability in the sampling distribution for two independent samples is more complex and must take into account variability coming from both samples. Hence the more complex formula for the standard error.

$$SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Can directed reading activities in the classroom help improve reading ability? A class of 21 third-graders participates in these activities for 8 weeks while a control classroom of 23 third-graders follows the same curriculum without the activities. After 8 weeks, all children take a reading test (scores in table).

Treatment group				Control group			
24	61	59	46	42	33	46	37
43	44	52	43	43	41	10	42
58	67	62	57	55	19	17	55
71	49	54		26	54	60	28
43	53	57		62	20	53	48
49	56	33		37	85	42	

Group	n	$\overline{x}$	s
Treatment	21	51.48	11.01
Control	23	41.52	17.15

95% confidence interval for ($\mu_1 - \mu_2$), with df = 20 conservatively → t^* = 2.086:

$$CI: \ (\bar{x}_1 - \bar{x}_2) \pm m; \quad m = t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = 2.086 * 4.31 \approx 8.99$$

With 95% confidence, ($\mu_1 - \mu_2$), falls within 9.96 ± 8.99 or 1.0 to 18.9.

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
	50%	60%	70%	80%	90%	95%	96%	98%	99%	99.5%	99.8%	99.9%
						Confidence level C						

Details of the two sample *t* procedures

The **true value of the degrees of freedom** for a two-sample *t*-distribution is quite lengthy to calculate. That's why we use an approximate value, df = smallest($n_1 - 1$, $n_2 - 1$), which errs on the conservative side (often smaller than the exact).

Computer software, though, gives the exact degrees of freedom—or the rounded value—for your sample data.

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

95% confidence interval for the reading ability study using the more precise degrees of freedom:

$$df = \frac{\left(\dfrac{11.01^2}{21} + \dfrac{17.15^2}{23}\right)^2}{\dfrac{1}{20}\left(\dfrac{11.01^2}{21}\right)^2 + \dfrac{1}{22}\left(\dfrac{17.15^2}{23}\right)^2}$$

$$= \frac{344.486}{9.099} = 37.86$$

	0.683	0.854	1.055	1.310	1.697	2.042	2.147	2.457
30	0.683	0.854	1.055	1.310	1.697	2.042	2.147	2.457
40	0.681	0.851	1.050	1.303	1.684	2.021	2.123	2.423
	50%	60%	70%	80%	90%	95%	96%	98%

Table C — Confidence level C

$$m = t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$m = 2.024 * 4.31 \approx 8.72$$

t-Test: Two-Sample Assuming Unequal Variances   **Excel**

	Treatment group	Control group
Mean	51.476	41.522
Variance	121.162	294.079
Observations	21	23
Hypothesized Mean Difference	-	
df	38	
t Stat	2.311	
P(T<=t) one-tail	0.013	
t Critical one-tail	1.686	
P(T<=t) two-tail	0.026	
t Critical two-tail	2.024	*t**

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means					95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Reading Score	Equal variances assumed	2.362	.132	2.267	42	.029	9.95445	4.39189	1.09125	18.81765
	Equal variances not assumed			2.311	37.855	.026	9.95445	4.30763	1.23302	18.67588

Pooled two-sample procedures

There are two versions of the two-sample *t*-test: one **assuming equal variance ("pooled 2-sample test")** and one **not assuming equal variance ("unequal" variance,** as we have studied) for the two populations. They have slightly different formulas and degrees of freedom.
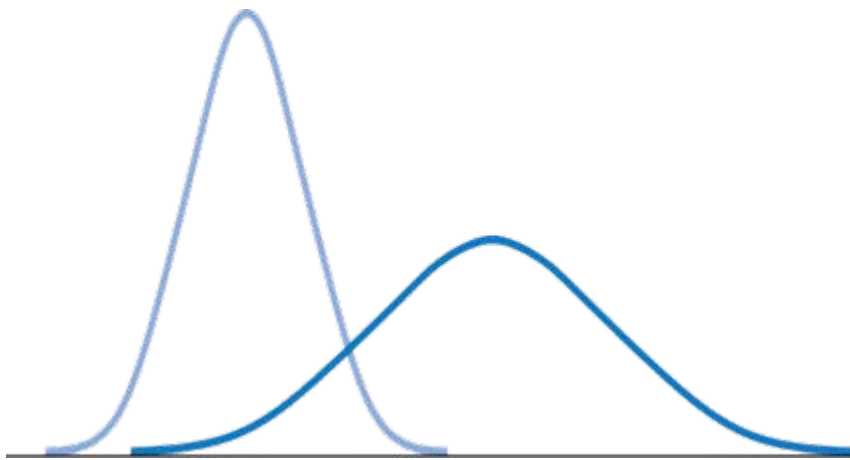
Two normally distributed populations with unequal variances

The pooled (equal variance) two-sample *t*-test was often used before computers because it has exactly the *t* distribution for degrees of freedom $n_1 + n_2 - 2$.

However, the assumption of equal variance is hard to check, and thus the unequal variance test is safer.

When both population have the *same* standard deviation, the **pooled estimator of σ²** is:

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

The sampling distribution for $(x_1 - x_2)$ has exactly the *t* distribution with **$(n_1 + n_2 - 2)$ degrees of freedom**.

A level *C* confidence interval for $\mu_1 - \mu_2$ is (with area C between $-t^*$ and t^*)

$$(\overline{x}_1 - \overline{x}_2) \pm t^* s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$t = \frac{\overline{x}_1 - \overline{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

To test the hypothesis $H_0: \mu_1 = \mu_2$ against a one-sided or a two-sided alternative, compute the pooled two-sample *t* statistic for the $t(n_1 + n_2 - 2)$ distribution.

Which type of test? One sample, paired samples, two samples?

- Comparing vitamin content of bread immediately after baking vs. 3 days later (the same loaves are used on day one and 3 days later).

- Comparing vitamin content of bread immediately after baking vs. 3 days later (tests made on independent loaves).

- Average fuel efficiency for 2005 vehicles is 21 miles per gallon. Is average fuel efficiency higher in the new generation "green vehicles"?

- Is blood pressure altered by use of an oral contraceptive? Comparing a group of women not using an oral contraceptive with a group taking it.

- Review insurance records for dollar amount paid after fire damage in houses equipped with a fire extinguisher vs. houses without one. Was there a difference in the average dollar amount paid?

Section 8.2

- Comparing Two Proportions:
- ✓ tests of significance for difference in proportions
- ✓ confidence intervals for difference in proportions
- ✓ relative risk

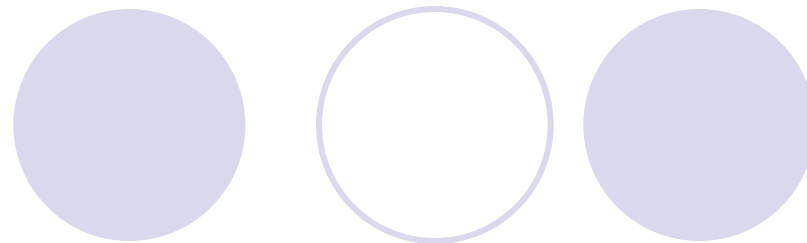# Comparing Two Proportions

● Notation:

Population	Population proportion	Sample size	Count of successes
1	p_1	n_1	X_1
2	p_2	n_2	X_2

Comparing Two Proportions

- **SRS of size n_1** from a large population having proportion p_1 of successes and independent **SRS of size n_2** from another large population having proportion p_2 of successes.

- $\hat{p}_1$ is an estimator of p_1

- $\hat{p}_2$ is an estimator of p_2:

$$\hat{p}_1 = \frac{X_1}{n_1}, \quad \hat{p}_2 = \frac{X_2}{n_2}$$

We have

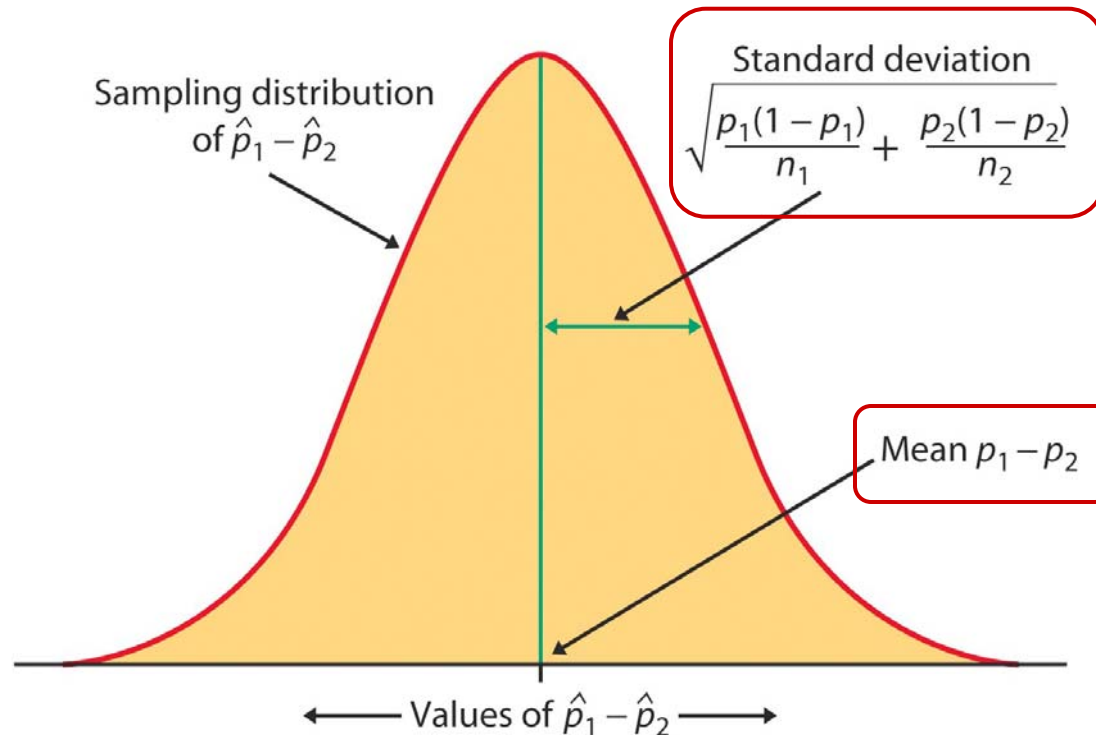$$\mu_{\hat{p}_1} = p_1 \qquad \sigma_{\hat{p}_1} = \sqrt{\frac{p_1(1-p_1)}{n_1}}$$

$$\mu_{\hat{p}_2} = p_2 \qquad \sigma_{\hat{p}_2} = \sqrt{\frac{p_2(1-p_2)}{n_2}}$$

Comparing two independent samples

We often need to compare two treatments used on **independent** samples. We can compute the difference between the two sample proportions and compare it to the corresponding, approximately normal sampling distribution for $(\hat{p}_1 - \hat{p}_2)$:

Sampling distribution of $\hat{p}_1 - \hat{p}_2$

Standard deviation
$$\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

Mean $p_1 - p_2$

Values of $\hat{p}_1 - \hat{p}_2$

Large-sample CI for two proportions

For two independent SRSs of sizes n_1 and n_2 with sample proportion of successes $\hat{p}_1$ and $\hat{p}_2$ respectively, an **approximate level *C* confidence interval** for $p_1 - p_2$ is

$$(\hat{p}_1 - \hat{p}_2) \pm m, \quad m \text{ is the margin of error}$$

$$m = z * SE_{diff} = z * \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}$$

C is the area under the standard normal curve between −z* and z*.

Use this method only when the populations are at least 10 times larger than the samples and the number of successes and the number of failures are each at least 10 in each samples.

Cholesterol and heart attacks

How much does the cholesterol-lowering drug Gemfibrozil help reduce the risk of heart attack? We compare the incidence of heart attack over a 5-year period for two random samples of middle-aged men taking either the drug or a placebo.

Standard error of the difference $p_1 - p_2$:

	H. attack	n	$\hat{p}$
Drug	56	2051	2.73%
Placebo	84	2030	4.14%

$$SE = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

$$SE = \sqrt{\frac{0.0273(0.9727)}{2051} + \frac{0.0414(0.9586)}{2030}} = 0.00764$$

The confidence interval is $(\hat{p}_1 - \hat{p}_2) \pm z * SE$

So the 90% CI is $(0.0414 - 0.0273) \pm 1.645*0.00746 = 0.0141 \pm 0.0125$

We are 90% confident that the percentage of middle-aged men who suffer a heart attack is 0.16% to 2.7% lower when taking the cholesterol-lowering drug.

Example - Exercise:

- In a highly-publicized study, doctors confirmed earlier observations that aspirin seems to help prevent heart attacks.

- The research project employed 21,996 male American physicians.

- Half of these took an aspirin tablet every other day, while the other half took a placebo on the same schedule. After 3 years, researchers determined that 139 of those who took aspirin and 239 of those who took placebo had had heart attacks.

- Determine whether these results indicate that aspirin is effective in reducing the incidence of heart attacks at significance level 0.05.

Solution:
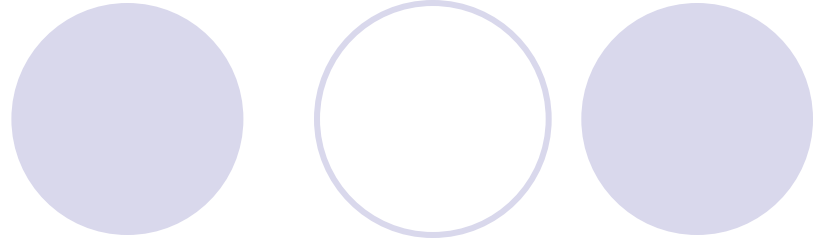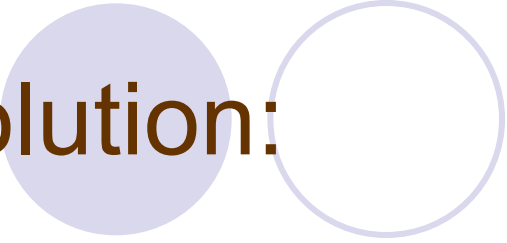
"Plus four" CI for two proportions

The "plus four" method again produces more accurate confidence intervals. We act as if we had four additional observations: one success and one failure in each of the two samples. The new combined sample size is $n_1 + n_2 + 4$ and the proportions of successes are:

$$\tilde{p}_1 = \frac{X_1 + 1}{n_1 + 2} \quad \text{and} \quad \tilde{p}_2 = \frac{X_2 + 1}{n_2 + 2}$$

An approximate level C confidence interval is:

$$CI: \quad (\tilde{p}_1 - \tilde{p}_2) \pm z^* \sqrt{\frac{\tilde{p}_1(1 - \tilde{p}_1)}{n_1 + 2} + \frac{\tilde{p}_2(1 - \tilde{p}_2)}{n_2 + 2}}$$

Use this when C is at least 90% and both sample sizes are at least 5.

Cholesterol and heart attacks

Let's now calculate the "plus four" CI for the difference in percentage of middle-aged men who suffer a heart attack (placebo – drug).

	H. attack	n	$\tilde{p}$
Drug	56	2051	2.78%
Placebo	84	2030	4.18%

$$\tilde{p}_1 = \frac{X_1 + 1}{n_1 + 2} = \frac{56 + 1}{2051 + 2} \approx 0.0278 \quad \text{and} \quad \tilde{p}_2 = \frac{X_2 + 1}{n_2 + 2} = \frac{84 + 1}{2030 + 2} \approx 0.0418$$

Standard error of the population difference $p_1 - p_2$:

$$SE = \sqrt{\frac{\tilde{p}_1(1 - \tilde{p}_1)}{n_1 + 2} + \frac{\tilde{p}_2(1 - \tilde{p}_2)}{n_2 + 2}} = \sqrt{\frac{0.0278(0.9722)}{2053} + \frac{0.0418(0.9582)}{2032}} = 0.0057$$

The confidence interval is $(\tilde{p}_1 - \tilde{p}_2) \pm z * SE$

So the 90% CI is (0.0418 − 0.0278) ± 1.645*0.00573 = 0.014 ± 0.0094

We are 90% confident that the percentage of middle-aged men who suffer a heart attack is 0.46% to 2.34% lower when taking the cholesterol-lowering drug.

Test of significance

If the null hypothesis is true, then we can rely on the properties of the sampling distribution to estimate the probability of drawing 2 samples with proportions $\hat{p}_1$ and $\hat{p}_2$ at random.
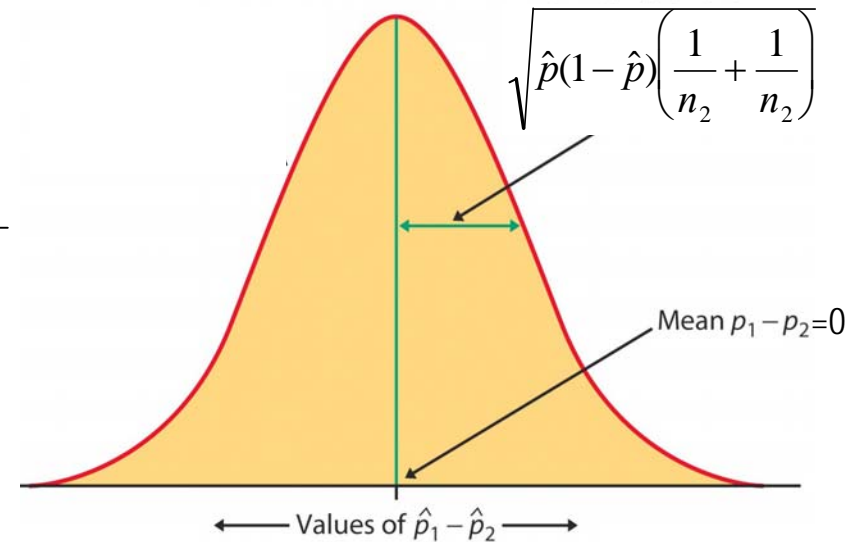
$H_0 : p_1 = p_2 = p$

Our best estimate of p is $\hat{p}$,

the pooled sample proportion

$$\hat{p} = \frac{\text{total successes}}{\text{total observations}} = \frac{\text{count}_1 + \text{count}_2}{n_1 + n_2}$$

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\dfrac{1}{n_2} + \dfrac{1}{n_2}\right)}}$$

$$\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_2} + \frac{1}{n_2}\right)}$$

Mean $p_1 - p_2 = 0$

Values of $\hat{p}_1 - \hat{p}_2$

This test is appropriate when the populations are at least 10 times as large as the samples and all counts are at least 5 (number of successes and number of failures in each sample).

Gastric Freezing

Gastric freezing was once a treatment for ulcers. Patients would swallow a deflated balloon with tubes, and a cold liquid would be pumped for an hour to cool the stomach and reduce acid production, thus relieving ulcer pain. **The treatment was shown to be safe, significantly reducing ulcer pain**, and so widely used for years.
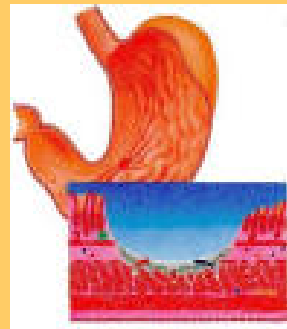
A randomized comparative experiment later compared the outcome of gastric freezing with that of a placebo: 28 of the 82 patients subjected to gastric freezing improved, while 30 of the 78 in the control group improved.

$$H_0: p_{gf} = p_{placebo} \qquad\qquad \hat{p}_{pooled} = \frac{28+30}{82+78} = 0.3625$$
$$H_a: p_{gf} > p_{placebo}$$

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\dfrac{1}{n_1}+\dfrac{1}{n_2}\right)}} = \frac{0.341-0.385}{\sqrt{0.363*0.637\left(\dfrac{1}{82}+\dfrac{1}{78}\right)}} = \frac{-0.044}{\sqrt{0.231*0.025}} = -0.499$$

Conclusion: The gastric freezing was no better than a placebo (p-value 0.69), and this treatment was abandoned. **ALWAYS USE A CONTROL!**

Relative risk

Another way to compare two proportions is to study the ratio of the two proportions, which is often called the **relative risk (RR).** A relative risk of 1 means that the two proportions are equal.

The procedure for calculating confidence intervals for relative risk is more complicated (use software) but still based on the same principles that we have studied.

The age at which a woman gets her first child may be an important factor in the risk of later developing breast cancer. An international study selected women with at least one birth and recorded if they had breast cancer or not and whether they had their first child before their 30th birthday or after.

	Birth age 30+	Sample size	$\hat{p}$
Cancer	683	3220	21.2%
No	1498	10,245	14.6%

$$RR = \frac{.212}{.146} \approx 1.45$$

Women with a late first child have 1.45 times the risk of developing breast cancer.