

Recall that **statistical inference** is inferring information about a population from information about a sample. We're generally talking about one of two things:

1. Estimating parameters (confidence intervals)
2. Hypothesis testing (significance tests)

The same basic ideas that we used when computing confidence intervals and evaluating hypothesis tests for means of a quantitative variable can be applied in a number of related situations. The best way to think of these different situations is as variations on an inference theme. To make this easier, we will use a systematic notation scheme throughout:

- parameters (population)
  - $p$ , proportion (of a categorical variable)
  - $\mu$ , mean (of quantitative variable)
  - $\sigma$ , standard deviation
- statistics (sample)
  - $n$ , sample size
  - $X$ , count (of a categorical variable)
  - $\hat{p} = \frac{X}{n}$ , proportion (of a categorical variable)
  - $\bar{x}$ , mean (of quantitative variable)
  - $s$ , standard deviation
- sampling distribution
  - $SD$ , standard deviation of the sampling distribution ( $\sigma_{\bar{x}}$  or  $\sigma_{\hat{p}}$  are also used for this)
  - $SE$ , standard error of the sampling distribution (an estimate for  $SD$ )
  - $\mu_{\hat{p}}, \mu_{\bar{x}}$ , mean of sampling procedure (for when determining  $\hat{p}$  and  $\bar{x}$ , respectively)

Subscripts will be used to indicate MULTIPLE POPULATIONS/SAMPLES

The procedures involving the  $z$  (normal) and  $t$  distributions are all very similar.

- To do a **hypothesis test**, compute

$$t \text{ or } z = \frac{\text{data value} - \text{hypothesis value}}{SD \text{ or } SE},$$

and compare with the appropriate distribution (using tables A and D).

- To compute a **confidence interval**, first determine the critical value for the desired level of confidence ( $z^*$  or  $t^*$ ), then the confidence interval is

$$\text{data value} \pm (\text{critical value})(SD \text{ or } SE).$$

## 1 Matched Pairs

Inference for matched pairs really is hardly new because generally we combine the values of two quantitative variables to form one new quantitative variable, and then we apply our inference procedures exactly as before. The most common way to combine the variables is simply take the difference between the two variables.

**Example.** Two varieties of oats were compared in an experiment to determine which variety had the higher yield. Since soil type also affects yield, the experimenter blocked out its effect by planting each variety of oats in seven different types of soil. With the data paired by soil types as given below, does it appear that variety A has the higher mean yield?

Soil type	Yield		x = A-B
	A	B	
1	71.2	65.2	6.0
2	72.6	60.7	11.9
3	47.8	42.8	5.0
4	76.9	73.0	3.9
5	42.5	41.7	0.8
6	49.6	56.6	-7.0
7	62.8	57.3	5.5

$$\text{mean}(x) = 3.7286 \quad \text{sd}(x) = 5.7792$$

$$H_0 : \text{MEAN DIFFERENCE } (A - B) \mu = 0$$

$$H_a : \text{MEAN DIFFERENCE } \mu > 0$$

$$SE = \frac{5.7792}{\sqrt{7}} = 2.1843$$

$$\text{p-value} = \text{BETWEEN } 0.025 \text{ AND } 0.05$$

$$95\% \text{ CI for difference: } (-1.613, 9.070)$$

**Example.** To test the effect of continuous music on factory workers' output, each of 7 workers was observed for one month with music and one month without. Given the following results, does music help? Find a 90% confidence interval for the mean difference in output.

	1	2	3	4	5	6	7
Average output with music	8.4	5.0	7.2	6.6	6.5	8.7	5.9
Average output w/o music	7.4	6.1	8.0	6.4	6.8	8.8	6.6
Diff	1.0	-1.1	-0.8	0.2	-0.3	-0.1	-0.7

$$\text{mean}(\text{Diff}) = -.257, \quad \text{sd}(\text{Diff}) = .709$$

ANS:  $(-0.778, 0.264)$

## 2 Comparing Two Means

A two-sample problem is one in which:

1. THE GOAL IS TO COMPARE THE RESPONSES IN TWO GROUPS
2. EACH GROUP IS CONSIDERED TO BE A SAMPLE FROM A DISTINCT POPULATION
3. THE RESPONSES IN EACH GROUP ARE INDEPENDENT FROM THOSE IN THE OTHER GROUP

The difference between two-sample problems and matched pairs problems is THAT IN MATCHED PAIRS WE HAVE TWO MEASUREMENTS THAT ARE DEPENDENT IN SOME WAY AND ONLY ONE POPULATION OF INTEREST WHEREAS IN TWO-SAMPLE PROBLEMS WE HAVE INDEPENDENT MEASUREMENTS FROM TWO DISTINCT POPULATIONS

So in a two-sample problem we want to compare  $\mu_1$  with  $\mu_2$ . We do this by drawing a sample from each population and calculating  $\bar{x}_1$  and  $\bar{x}_2$ . In order to know what  $\bar{x}_1$  and  $\bar{x}_2$  tell us about the difference between  $\mu_1$  and  $\mu_2$  we need to know about the sampling distribution for  $\bar{x}_1 - \bar{x}_2$ .

Assuming each population has a normal distribution with means  $\mu_i$  and standard deviations  $\sigma_i$ , we already know that

$$\bar{x}_1 \sim N(\mu_1, \sigma_1/\sqrt{n_1}) \quad \text{and} \quad \bar{x}_2 \sim N(\mu_2, \sigma_2/\sqrt{n_2})$$

Using our rules for combining means and variances, we see that

$$\bar{x}_1 - \bar{x}_2 \sim N(\mu_1 - \mu_2, SD), \text{ where } SD = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Of course, we won't usually know  $\sigma_1$  and  $\sigma_2$ , so we need to estimate  $SD$  using  $\underline{s}_1$  and  $\underline{s}_2$ . There are two ways to do this:

1. NOT ASSUMING THE TWO VARIANCES ARE EQUAL ( $\sigma_1^2 = \sigma_2^2$ )
2. ASSUMING THE TWO VARIANCES ARE EQUAL ( $\sigma_1^2 = \sigma_2^2$ )

## 2.1 Unequal variances

In this case we simply replace each  $\underline{\sigma}_i$  with  $\underline{s}_i$ , so

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

Unfortunately the  $t$  statistic computed from this does not have a  $t$ -distribution as we might have hoped. Why? A  $t$ -distribution replaces a  $N(0,1)$  distribution only when a single population standard deviation  $\sigma$  is replaced by a single sample standard deviation  $s$ . In this case, we replaced two standard deviations ( $\sigma_1$  and  $\sigma_2$ ) with their estimates ( $s_1$  and  $s_2$ ). Nevertheless, we can approximate the distribution with an **approximate degrees of freedom**. There are two ways to do this:

1. MESSY FORMULA
2. SMALLER OF  $df_1$  AND  $df_2$

Method 1 is typically used by computer software, but is a bit messy to be done by hand. Furthermore, the simpler method is *conservative*: the value it gives for  $df$  is always on the low side, which means that it will give CIs that are bit TOO WIDE and p-values that are a bit TOO BIG.

Now that we know the distribution involved and a value for  $SE$ , we are all set to do hypothesis testing or to compute CIs.

**Example.** An agronomist has developed a new plant food. She hopes it will improve yield. To find out, she treated 48 plants with the new food and obtained a mean yield of 24.4 lbs ( $s = 4.8$  lbs). 45 identical plants were untreated and had a mean yield of 22.3 lbs ( $s = 2.3$  lbs). Do the data provide sufficient evidence to determine that the new plant food is better than no treatment?

$$H_0 : \mu_{\text{TREATED}} - \mu_{\text{UNTREATED}} = 0$$

$$H_a : \mu_{\text{TREATED}} - \mu_{\text{UNTREATED}} > 0$$

$$t = \frac{(24.4 - 22.3) - 0}{\sqrt{\frac{4.8^2}{48} + \frac{2.3^2}{45}}} = 2.7166$$

WE REJECT THE NULL HYPOTHESIS. THE DATA IS SIGNIFICANT AT THE 5% LEVEL TO CONCLUDE THAT THE NEW PLANT FOOD IS BETTER THAN NO TREATMENT ( $t = 2.7166$ ,  $DF = 44$ ,  $0.0025 < P < 0.005$ ).

**Example.** The weather bureau measured the ozone level at 5 random locations in Orange City before a cold front moved through, and at 5 different random locations afterward. Test whether there is a significant drop in the ozone level after the front has moved through.

Time	$n$	mean	variance
Before front	5	0.122	.00067
After front	5	.094	.00016

$$H_0 : \text{DIFFERENCE IN MEANS (AFTER - BEFORE)} \mu_{\text{AFTER}} - \mu_{\text{BEFORE}} = 0$$

$$H_a : \text{DIFFERENCE IN MEANS (AFTER - BEFORE)} \mu_{\text{AFTER}} - \mu_{\text{BEFORE}} < 0$$

$$t = \frac{(0.094 - 0.122) - 0}{\sqrt{\frac{0.00067}{5} + \frac{0.00016}{5}}} = -2.1732$$

TO GET THE SINGLE-TAIL PROBABILITY, WE TAKE THE ABSOLUTE VALUE AND CONSULT TABLE D FOR  $P(t \geq 2.1732)$ . FOR  $df = 4$  WE GET  $0.025 < P < 0.05$ , SHOWING THAT THE DATA IS SIGNIFICANT AT THE 5% LEVEL FOR REJECTING  $H_0$ , BUT NOT AT THE 1% LEVEL.

How would the design of this study need to be changed to make it a matched pairs situation? TAKE THE BEFORE AND AFTER READINGS AT THE SAME LOCATION.

**Example.** The scores of two groups of prison inmates on a rehabilitation test are summarized below:

	First offenders	Repeat offenders
Mean	300	305
Variance	20	4
Sample size	16	13

Compute a 95% confidence interval for the difference between average rehabilitation scores for repeat and first offenders. Is there evidence significant at the 5% level that the scores for the two groups are different?

$$\bar{x}_2 - \bar{x}_1 \pm t^*(SE) = 5 \pm 2.179 \left( \sqrt{\frac{20}{16} + \frac{4}{13}} \right) \text{ OR } (2.280, 7.720).$$

WE MAY USE THIS CI IN PLACE OF A SEPARATE TEST OF SIGNIFICANCE, SINCE IT WAS A 2-SIDED ALTERNATIVE HYPOTHESIS AND ALL WE ARE ASKED IS WHETHER THE DATA IS SIGNIFICANT AT THE 5% LEVEL. (WE'RE NOT ASKED FOR THE  $P$ -VALUE.) IT IS SIGNIFICANT AT THE 5% LEVEL, SINCE 0 (THE HYPOTHESIZED DIFFERENCE IN  $H_0$ ) IS NOT IN OUR 95% CI. WE WOULD REJECT  $H_0$  AND CONCLUDE THAT THE MEAN REHABILITATION SCORE FOR THE TWO GROUPS ARE DIFFERENT.

## 2.2 Equal variances: pooled two-sample procedures

If the two population distributions have the same variance, then we do have a situation that is exactly a  $t$  distribution, and we *pool* data from both samples to estimate the common variance.

We will **use the pooled estimate** if we have reason to believe that the two population variances are equal. One common **rule of thumb** is that if the variances of the two samples are within a factor of 4 (so  $s_1$  and  $s_2$  will be within a factor of  $\underline{2}$ ), then we can use the pooled estimate. Computer software will usually have clearly marked options for the two types of two sample tests or else print out both results and leave it to you to decide which should be used.

When we do this, our pooled estimate for the variance in each population is

$$s_p = \sqrt{\frac{(df_1)s_1^2 + (df_2)s_2^2}{df_1 + df_2}}$$

and

$$SE = \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}} = s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

The degrees of freedom of the  $t$  distribution in this case is the sum of the degrees of freedom from the two samples:

$$df = df_1 + df_2 = (n_1 - 1) + (n_2 - 1) = n_1 + n_2 - 2$$

Once again, with the distribution in hand, we are all set to do hypothesis testing or to compute CIs.

## 2.3 Robustness

A test is considered *robust* if the probability calculations required are insensitive to violations of the assumptions made. Just as with 1-sample  $t$  procedures, the 2-sample  $t$  procedures are based upon the assumption that the two samples come from populations which are normal. On pp. 515–517, the authors of our text explain that the 1-sample  $t$  procedures are fairly robust, giving specifics. In fact, the two-sample  $t$  procedures are even more robust than the one-sample  $t$  procedures.

When the sizes of the two samples are equal and the distributions of the two populations being compared have similar shapes, p-values from Table D are quite accurate for a broad range of distributions when the sample sizes are as small as  $n_1 = n_2 = 5$ .

When the two population distributions have different shapes, larger samples are needed.

The guidelines we gave before (sample sizes  $< 15$ ,  $15 - 40$ ,  $\geq 40$ ) can be adapted to two-sample procedures by replacing the one-sample sample size with the sum of the sample sizes  $n_1 + n_2$ .

### 3 Categorical Data

#### 3.1 One-proportion procedures

We already know

1. the sampling distribution for  $\hat{p}$  is based on the BINOMIAL distribution,
2. the sampling distribution for  $\hat{p}$  can be approximated by a NORMAL distribution provided  $np \geq 10$  AND  $nq \geq 10$
3.  $SD_{\hat{p}} = \frac{\sqrt{p(1-p)}}{\sqrt{n}}$

##### 3.1.1 Hypothesis testing

This is all we need to know for **hypothesis testing**, since the hypothesis provides us with a value for  $p$ , from which we can get the value of  $SD$ .

**Example.** A pollster contacts 225 people in Statville and asks if they plan to vote for or against Referendum A. 99 say they will vote in favor, 125 say they will vote against. Is this sufficient evidence to predict the outcome of the vote?

$$H_0 : p = 0.5, H_a : p \neq 0.5$$

$$z = \frac{99/224 - 0.5}{\sqrt{\frac{(0.5)(0.5)}{224}}} \approx -1.7372.$$

TO GET THE TWO-TAILED PROBABILITY, WE ADD UP  $P(Z < -1.74) + P(z > 1.74) \approx 2(0.0409) = 0.0818$ . SO, EVEN THOUGH OUR SAMPLE HAS JUST  $99/224 = 44.2\%$  OF PEOPLE VOTING YES, IT WILL HAPPEN ABOUT 8% OF THE TIME THAT A SAMPLE PROPORTION IS THIS FAR AWAY FROM 50% EVEN THOUGH  $p = 0.5$ . ONLY AT THE 10% LEVEL WOULD WE REJECT THE NULL HYPOTHESIS, AND CONCLUDE THAT OUR SAMPLE WAS SIGNIFICANT TO CONCLUDE THE OUTCOME OF THE REFERENDUM.

##### 3.1.2 Confidence intervals

In order to compute **confidence intervals** we need to estimate  $SD$ , since we don't know what  $p$  is. Our estimate will come by replacing  $p$  with  $\hat{p}$ . Once we do that we get

$$SE = \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}$$

The distribution is still approximately normal (not a  $t$  distribution because the standard deviation is not an independent parameter in this case, it is determined by  $p$ ). So we use the normal distribution (Table A) for confidence intervals, too.

**Example.** The National Transportation Safety Board conducted a study of truck drivers killed in highway accidents. They found that 24 of 185 drivers tested positive for alcohol. Obtain a 95% confidence interval for the true percentage of truck driver deaths in which the truck driver had a positive level of alcohol.

$$\hat{p} \pm 1.96 \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \frac{24}{185} \pm 1.96 \sqrt{\frac{(0.1297)(0.8703)}{185}}, \text{ OR } (0.081, 0.178).$$

If they were reporting this on the news, they would say: It is estimated that 13% of truck drivers killed in highway accidents tested positive for alcohol (margin of error  $\pm 5$  percentage points).

**Example.** A special Newsweek on aging in fall/winter 2001 reported the results of a national survey of U.S. adults. They reported that 20% of women and 6% of men would like a face lift. At the bottom of the percentages, they wrote, "For this special Newsweek poll, Princeton Survey Research Associates interviewed a random national sample of 801 adults 45-65 years old by telephone July 13-17. The margin of error is  $\pm 5$  percentage points for women,  $\pm 6$  for men".

So, what does this mean?

THIS MEANS THAT THE TRUE PROPORTION OF WOMEN WHO WOULD LIKE A FACE LIFT LIKELY LIES BETWEEN 10% AND 20%, WHILE THE TRUE PROPORTION FOR MEN LIKELY LIES BETWEEN 0% AND 12%. (THESE ARE PROBABLY MARGINS OF ERROR ASSOCIATED WITH 95% CONFIDENCE.

### 3.1.3 Determine sample size

Since we don't know either  $p$  or  $\hat{p}$  when we decide how large a sample to get, we do one of the following:

- MAKE AN EDUCATED GUESS AT  $p$ ,
- use the fact that  $SD$  is largest when  $p = .5$ , and use that value.

**Example.** In order to get a margin of error of 6%, how many people must be surveyed in a public opinion poll? (Use 95% confidence level.)

$$0.06 \geq 1.96 \left( \frac{0.5}{\sqrt{n}} \right) \Rightarrow n \geq \left( \frac{(1.96)(0.5)}{0.06} \right)^2 = 266.78.$$

SO, TAKE A SAMPLE SIZE  $n$  OF 267 OR LARGER.

**Example.** A flower bulb company claims that 95% of their bulbs will come up. To get a 99% confidence interval that with a margin of error that is at most  $\pm 2\%$ , how many bulbs must be planted? Is it OK to plant them all in my front yard?

$$0.02 \geq 2.576 \sqrt{\frac{(0.95)(0.05)}{n}} \Rightarrow n \geq \frac{(2.576)^2(0.95)(0.05)}{(0.02)^2} = 788.00.$$

WE WOULD NEED TO PLANT AT LEAST 788 BULBS. IT WOULD NOT BE OK TO PLANT THEM ALL IN MY FRONT YARD, AS COMPETITION FOR ROOT SPACE AND RESOURCES WOULD BECOME A CONFOUNDING VARIABLE.

## 3.2 Comparing two proportions

To compare two proportions, we need to know the sampling distribution of  $\hat{p}_1 - \hat{p}_2$  (the difference between the two proportions). Just as for comparing two means, we can use our rules for means and variances to determine the mean and variance of the sampling distribution for  $\hat{p}_1 - \hat{p}_2$

$$\text{mean} = p_1 - p_2 \qquad SD = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_1}}$$

### 3.2.1 Confidence Intervals

When we compute confidence intervals, of course, we we don't know  $p$  but we can estimate  $SD$  by

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

The resulting distribution is still approximately normal.

**Example.** In a recent survey, about  $2/3$  of the men and half the women classified themselves as aggressive drivers. Assume that 1200 men and 1200 women were interviewed, and that the sample can be regarded as representative of all men and women drivers. Estimate with 95% confidence the difference in the proportions of men and women who think they drive aggressively.

$$\left(\frac{1}{2} - \frac{1}{3}\right) \pm 1.96 \sqrt{\frac{(0.667)(0.333)}{1200} + \frac{(0.5)(0.5)}{1200}}, \text{ OR } (0.128, 0.206)$$

### 3.2.2 Hypothesis testing

The most common null hypothesis for comparing two proportions is

$$H_0: p_1 - p_2 = 0$$

If  $H_0$  is true, then  $p_1 = p_2$ . Let's call this common proportion  $p$ . Our best (pooled) estimate for  $p$  is then

$$\hat{p} = \frac{X_1 + X_2}{n_1 + n_2}$$

and

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

Once again, the distribution of  $\hat{p}_1 - \hat{p}_2$  will be approximately normal, so we will use Table A.

#### Example.

A study indicated that a drug designed to rid animals of worms greatly improves the chances of survival for advanced Stage C colon cancer patients. The drug (levamisole) was combined with a standard cancer medicine (fluorouracil) to treat 304 patients, and 103 suffered relapses. A control group of 315 Stage C colon cancer patients had 155 relapses. Test at the 0.05 level to see if the recurrence rate for the treatment group is significantly lower than the recurrence rate for the control group.

$$H_0: p_{\text{TR}} - p_{\text{CTRL}} = 0, \quad H_a: p_{\text{TR}} - p_{\text{CTRL}} < 0.$$

$$\hat{p} = \frac{103 + 155}{304 + 315} \approx 0.417.$$

$$z = \frac{(0.3388 - 0.4921) - 0}{\sqrt{(0.417)(0.583) \left(\frac{1}{304} + \frac{1}{315}\right)}} \approx -3.87.$$

THE ONE-TAILED PROBABILITY  $P(Z \leq -3.87) < 0.0003$ . THIS IS SIGNIFICANT AT THE 5% LEVEL, SO WE ACCEPT THE ALTERNATIVE HYPOTHESIS THAT THE TREATMENT GROUP HAS A LOWER INCIDENCE OF RECURRENCE.

### 3.2.3 Relative Risk

Relative Risk (RR) = THE RATIO OF TWO PROPORTIONS

If  $RR = 1$ , THE TWO PROPORTIONS ARE EQUAL

If  $RR \neq 1$ : THE TWO PROPORTIONS ARE NOT EQUAL

The RR is commonly used in the study of epidemiology (the study of patterns of disease occurrence in human populations and of the factors that influence these patterns).

**Example.** A client was studying the rate of clinical depression in those who had been abused in some way. Of the 44 who had been abused, 23 suffered from depression. Of the 37 in her study who had not been abused, 10 suffered from depression. What is the relative risk?

Software can also compute confidence intervals for relative risk.