1.

Answer Key original

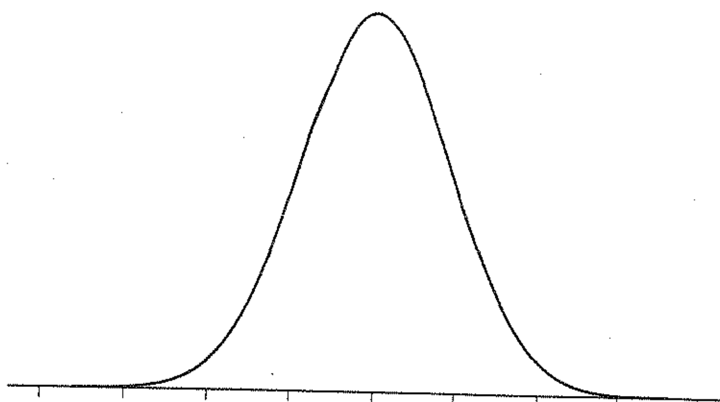
Solution Activity 17-5 Baseball "Bing Bang"

a. The null hypothesis is that the proportion of all major-league baseball games that contain a big bang is three-fourths. In symbols, the null hypothesis is H_0: $\pi = .75$.

b. The alternative hypothesis is that less than three-fourths of all major-league baseball games contain a big bang. In symbols, the alternative hypothesis is H_a: $\pi < .75$.

c. The CLT applies here because $95(.75) = 71.25$ is greater than 10 and $95(.25) = 23.75$ is also greater than 10. According to the CLT, the sample proportion would vary approximately normally with mean .75 and standard deviation

$$\sqrt{\frac{(.75)(.25)}{95}} \approx .0444$$



d. The sample proportion of games in which a big bang occurred is

$$\hat{p} = \frac{47}{95} \approx .495$$

e. Yes, this sample proportion is less than .75, as Marilyn conjectured.

f. The test statistic is

$$z = \frac{.495 - .75}{\sqrt{\frac{(.75)(.25)}{95}}} \approx \frac{.495 - .75}{.0444} \approx -5.74$$

This statistic says that the observed sample result is almost six standard deviations below what the grandfather conjectured. This z-score is way off the chart in Table II, indicating that the p-value is virtually zero.

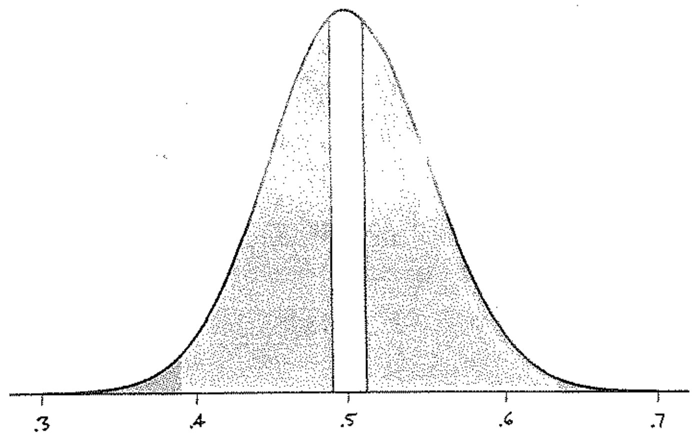
g. Yes, this very small p-value indicates that the sample data provide extremely strong evidence against the grandfather's claim. There is extremely strong evidence that less than 75% of all major-league baseball games contain a big bang. The null (grandfather's) hypothesis would be rejected at the $\alpha = .01$ level.

h. The hypotheses for testing Marilyn's claim are H_0: $\pi = .5$ vs. H_a: $\pi \neq .5$.

i. The test statistic is

$$z = \frac{.495 - .5}{\sqrt{\frac{(.5)(.5)}{95}}} \approx \frac{.495 - .5}{.0513} \approx -0.10$$

The p-value is $2(.4602) = .9204$.



j. This p-value is not small at all, suggesting that the sample data are quite consistent with Marilyn's hypothesis that half of all games contain a big bang. The sample data provide no reason to doubt Marilyn's hypothesis.

k. A 95% confidence interval for π (the population proportion of games that contain a big bang) is given by

$$\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

which is

$$.495 \pm 1.96 \sqrt{\frac{(.495)(.505)}{95}}$$

which is $.495 \pm .101$, which is the interval from .394 to .596. Therefore, you are 95% confident that between 39.4% and 59.6% of all major-league baseball games contain a big bang. The grandfather's claim (75%) is not within this interval or even close to it, which explains why it was so soundly rejected. Marilyn's conjecture (50%) is well within this interval of plausible values, which is consistent with it not being rejected.

Watch Out

- Remember that the hypotheses are always statements about a parameter, not about a statistic. The whole point is to see what you can infer about an unknown parameter value based on a sample statistic.
- Remember to use the hypothesized value of the parameter (denoted by π_0) of the test statistic calculation under the square root in the denominator and also in checking the technical conditions. It is easy to mistakenly use the sample proportion $\hat{p}$ in those calculations.
- Try to carry many decimal places of accuracy in intermediate calculations. If you round too much in an early calculation, that error can get magnified in subsequent calculations.
- Again remember that you do not "accept" a null hypothesis, even one with a p-value as great as Marilyn's. The sample data are in very close agreement with Marilyn's hypothesis, but you still should not conclude that exactly 50% of all games contain a big bang.
- Even with an extremely small p-value, stop short of saying that the data *prove* that the null hypothesis is false. Even though you have overwhelming evidence against the grandfather's claim, you have not technically *proven* that his claim is wrong.

2. **Solution** Children's Television Viewing

The observational units are third- and fourth-grade students. The sample consists of the 198 students at two schools in San Jose. The population could be considered all American third- and fourth-graders, but it might be more reasonable to restrict the population to be all third- and fourth-graders in the San Jose area at the time the study was conducted.

The variable measured here is the *amount of television the student watches in a typical week,* which is quantitative. The parameter is the mean number of hours of television watched per week among the population of all third- and fourth-graders. This population mean is denoted by μ. The question asked about watching an average of two hours of television per day, so convert that to be 14 hours per week.

The null hypothesis is that third-and fourth-graders in the population watch an average of 14 hours of television per week (H_0: $\mu = 14$). The alternative hypothesis is that these children watch more than 14 hours of television per week on average (H_a: $\mu > 14$).

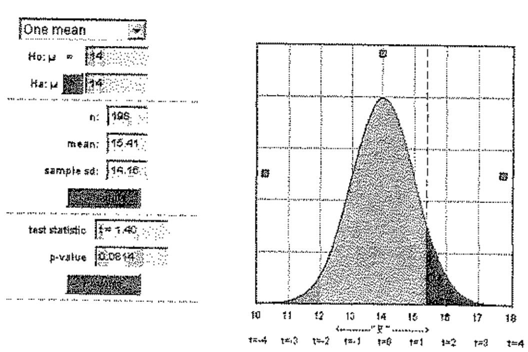
You should check the technical conditions for the t-test before you proceed:

- The sample of children was not chosen randomly; they all came from two schools in San Jose. You might still consider these children to be representative of third- and fourth-graders in San Jose, but you might not be willing to generalize to a broader population.
- The sample size is large enough (198 is far greater than 30) that the second condition holds regardless of whether or not the data on television watching follow a normal distribution. You do not have access to the child-by-child data in this case, so you cannot examine graphical displays; however, the large sample size assures you that the t-test is nevertheless valid to employ.

The sample size is $n = 198$; the sample mean is $\bar{x} = 15.41$ hours; and the sample standard deviation is $s = 14.16$ hours. The test statistic is

$$t = \frac{15.41 - 14}{14.16/\sqrt{198}} \approx 1.401$$

indicating the observed sample mean lies 1.401 standard errors above the conjectured value for the population mean. Looking in Table III, using the 100 df line (rounded down from the actual df of $198 - 1 = 197$), reveals the p-value (probability to the right of $t = 1.401$) to be between .05 and .10. Technology calculates the p-value more exactly to be .081.



This p-value is not less than the .05 significance level. The sample data, therefore, do not provide sufficient evidence to conclude that the population mean is greater than 14 hours of television watching per week. This conclusion stems from realizing that obtaining a sample mean of 15.41 hours or greater would not be terribly uncommon when the population mean is really 14 hours per week. If you had used a greater significance level (such as .10), which requires less compelling evidence in order to reject a hypothesis, then you would have concluded that the population mean exceeds 14 hours per week.

Watch Out

- The first step is to decide whether your test is about a proportion or a mean. One way to judge is to ask whether the variable (as measured on the individual observational units) is categorical or quantitative. Also be on the lookout for more obvious clues, such as the word *average* appearing in the statement of this problem.
- Remember the hypotheses are about parameters, not statistics. Also remember the hypothesized value comes from the research question, not from the sample data.
- When a sample is not chosen randomly, be cautious about generalizing the study results to a larger population. Try instead to think about what population the sample might be representative of, and even then be wary.
- Do not forget to include the $\sqrt{n}$ factor in calculating the value of the test statistic.

3. *Solution* Nicotine Lozenge

a. The explanatory variable is the type of lozenge (nicotine or placebo). The response variable is whether the smoker successfully abstains from smoking for the year. Both variables are categorical and binary.

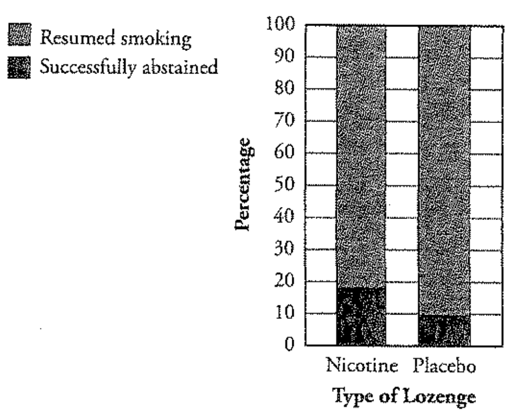
b. This is an experiment because the researchers *assigned* the subjects to take a particular kind of lozenge (nicotine or placebo).

c. The null hypothesis is that the nicotine lozenge is no more (or less) effective than the placebo, where effectiveness is measured by the proportion of smokers who successfully abstain from smoking for a year. The alternative hypothesis is that the nicotine lozenge is *more* effective than the placebo, meaning a higher proportion of smokers (who are interested in and might potentially use such a product) would successfully quit with a nicotine lozenge than with a placebo lozenge. In symbols, the hypotheses are H_0: $\pi_{nicotine} = \pi_{placebo}$ vs. H_a: $\pi_{nicotine} > \pi_{placebo}$, where π represents the population proportion of smokers who successfully abstain for a year if given the nicotine lozenge or the placebo.

d. The 2×2 table is shown here:

	Nicotine Lozenge	Placebo Lozenge	Total
Successfully Abstained	82	44	126
Resumed Smoking	377	414	791
Total	459	458	917

The following segmented bar graph shows those smokers taking the nicotine lozenge had a higher success rate (proportion) in this study than those smokers taking the placebo lozenge, almost twice as high (.179 vs. .096). But the graph also reveals that in both groups, many more smokers resumed smoking than were able to abstain successfully.



e. The sample proportions of smokers who successfully abstained from smoking in each group are

$$\hat{p}_{nicotine} = \frac{82}{459} \approx .179, \hat{p}_{placebo} = \frac{44}{458} \approx .096$$

The combined sample proportion of smokers who abstained is

$$\hat{p}_c = \frac{82 + 44}{459 + 458} = \frac{126}{917} \approx .137$$

The test statistic is

$$z = \frac{.179 - .096}{\sqrt{(.137)(1 - .137)\left(\frac{1}{459} + \frac{1}{458}\right)}} \approx 3.65$$

The p-value is the area to the right of 3.65 under the standard normal curve, which Table II reveals to be .0001. This test is valid because random assignment was used to put subjects in groups and because the sample size condition is also met: $459(.137) = 62.88$ is greater than 5, as are $458(.137) = 62.75$, $459(1 - .137) = 396.12$, and $458(1 - .137) = 395.25$.

This very small p-value indicates the experimental data provide very strong evidence against the null hypothesis, which means very strong evidence that the population proportion of smokers who would successfully abstain from smoking is higher with the nicotine lozenge than with the placebo lozenge. Because this was a randomized experiment, you can conclude the nicotine lozenge *causes* an increase in the proportion of smokers who would successfully abstain, as compared to the placebo group. You are not told how the subjects were selected for the study, but you do know that they were hoping to quit smoking, so this conclusion should be limited to smokers hoping to quit. They were probably chosen from a particular geographic area, so you might not want to generalize beyond that area, but it is hard to imagine why smokers in one area would respond differently to these lozenges than smokers in another area.

f. A 95% confidence interval for the difference $\pi_{\text{nicotine}} - \pi_{\text{placebo}}$ is

$$(.179 - .096) \pm 1.96\sqrt{\frac{(.179)(1 - .179)}{459} + \frac{(.096)(1 - .096)}{458}}$$

which is $.083 \pm 1.96(.023)$, which is $.083 \pm .044$, which is the interval from .039 through .127. The researchers are 95% confident that the proportion of smokers using a nicotine lozenge who successfully quit would be higher than the successful proportion of smokers using a placebo lozenge by somewhere between .039 and .127. This interval procedure is valid because of random assignment and because the number of successes and failures in both groups exceeds 5 (the smallest of these numbers is 44 successes in the placebo group) with the sampling issues cautioned about in part e.

g. Notice this question calls for a confidence interval for a single proportion: π_{nicotine}, so you need to use the procedure from Topic 16:

$$\hat{p}_{\text{nicotine}} \pm z^* \sqrt{\frac{\hat{p}_{\text{nicotine}}(1 - \hat{p}_{\text{nicotine}})}{n_{\text{nicotine}}}}$$

which is $.179 \pm 1.96(.018)$, which is $.179 \pm .035$, which is the interval from .144 through .214. The researchers can be 95% confident that if the population of all smokers who wanted to quit were to use the nicotine lozenge, the proportion who would successfully abstain from smoking for one year would be between .144 and .214. So, even though the experiment provides strong evidence that the nicotine lozenge works better than a placebo, most smokers (more than 3/4) would be unable to quit even with the nicotine lozenge.

Watch Out

- Be sure to plug in proportions (which must be between 0 and 1), not percentages or counts, in these calculations.
- Be especially careful about rounding errors with these calculations. During intermediate steps of the calculation, carry as many decimal places of accuracy as

4 **Solution** Got a Tip?

a. This is a randomized experiment, because the waitress randomly assigned the two-person parties to be told her name or not as part of her greeting. We do have to consider that the waitress was not blind to the treatment condition and may have subconsciously provided better service for the customers she expected to give her a larger tip.

b. The explanatory variable is whether the waitress included her name as part of her greeting; this variable is categorical and binary. The response variable is the amount of the tip, which is quantitative.

c. The null hypothesis is that there is no effect from the waitress using her name in her greeting to customers. In other words, the null hypothesis says that the population mean tip amount will be the same when she uses her name as when she does not. The alternative hypothesis is that using her name has a positive effect, that the population mean tip amount is greater when she uses her name than when she does not. In symbols: H_0: $\mu_{\text{name}} = \mu_{\text{no name}}$ vs. H_a: $\mu_{\text{name}} > \mu_{\text{no name}}$.

d. A Type I error means the waitress decides using her name helps when it really doesn't, so she would waste the minimal effort of giving her name and reap no benefit from it. A Type II error means the waitress decides using her name is not helpful even though it actually is helpful, so she would not bother to give customers her name and would lose out on that benefit. Because the cost of giving her name is minimal, losing out on potential tips is probably more of a concern.

e. The test statistic is

$$t = \frac{5.44 - 3.49}{\sqrt{\frac{(1.75)^2}{20} + \frac{(1.13)^2}{20}}} \approx \frac{1.95}{0.466} \approx 4.19$$

Looking in Table III (t-Distribution Critical Values) with 19 degrees of freedom, reveals that this test statistic is off the chart, so the p-value is less than .0005.

f. Because the p-value is less than .05, you reject the null hypothesis at the $\alpha = .05$ level. Indeed, you would also reject the null hypothesis at the $\alpha = .01$ and even at the $\alpha = .001$ levels. The data and experimental design provide very strong evidence that giving her name to customers as part of her greeting does lead to higher tips on average.

g. You do not have enough information to check the technical conditions thoroughly. You do know the parties were randomly assigned to one group or the other. But the sample sizes are not large, so you should check whether the tip data could reasonably have come from normal distributions; however, you have only the summary statistics, not the actual tip amounts from each party, so you cannot check this condition. You should ask the waitress to provide the actual party-by-party tip amounts to help you assess the shape of the distribution of tip amounts.

h. A 95% confidence interval for $\mu_{\text{name}} - \mu_{\text{no name}}$ is

$$(5.44 - 3.49) \pm 2.093\sqrt{\frac{(1.75)^2}{20} + \frac{(1.13)^2}{20}}$$

which is $1.95 \pm 2.093(0.466)$, which is 1.95 ± 0.97, which is the interval (0.98, 2.92). You can be 95% confident that the waitress would earn between \$0.98 and \$2.92 more per party with a \$23.21 check, on average, by giving her name as part of her greeting (assuming that the tip amounts are roughly normally distributed).

i. You can conclude a causal link between the waitress giving her name and receiving higher tips on average. Random assignment should have assured that the only difference between the groups was whether the party was given the waitress's name. Because the group who was told her name gave significantly higher tips on average (p-value < .0001), you can attribute that to being told her name in her greeting unless the waitress gave better service to customers to whom she gave her name. The confidence interval enables you to say more: that giving her name to customers increases the waitress' tips by an average of about 1-3 dollars per dining party at Sunday brunch in this restaurant. But you must be cautious about generalizing this result to other waitresses because only one waitress participated in this study. Even for this particular waitress, you should be cautious about generalizing the results to customers beyond those who partake of Sunday brunch at that particular Charley Brown's restaurant in southern California. You should also remember that these p-value and confidence interval calculations are only valid if the tip amounts roughly follow a normal distribution.

Watch Out

The cautions in this topic are not substantially new; however, we want to remind you of some good habits that are easy to forget:

- Be sure to choose the correct procedure in the first place. A good first step, as we've emphasized since Topic 1, is to identify the observational units and variables. In this study, the observational units are the dining parties; the explanatory variable is whether the waitress gives her name or not (categorical, binary); and the response variable is the amount of tip (quantitative). A two-sample t-procedure is appropriate in situations like this, when the explanatory variable is binary and the response variable is quantitative. Another way to think of this situation is that you want to compare two groups on a quantitative response variable.
- Remember the hypotheses are always about parameters, denoted by Greek letters. It's easy to forget and put symbols for sample statistics in the hypotheses. But you don't need to test whether the sample means are equal in the two groups, because you *know* the values of the sample mean tip amounts. The question is what you can *infer* about the unknown *population* means.
- Again be careful not to let rounding errors creep into your calculations. Also be aware your answers may differ very slightly if you use technology to conduct these tests and construct these intervals.
- Always remember to check technical conditions before taking a test or interval result seriously. Also notice that the second technical condition here is an either/or statement: the distributions do not have to be normal if the sample sizes are large.
- Remember the scope of conclusions, with regard to causation and generalizability, depends on how the study is conducted.
- Do not forget to relate your conclusions to the context of the study. Do not simply say "reject H_0" and leave it at that. Such a conclusion would not be very helpful to this waitress. Rather, say the data provide very strong evidence that giving her name to customers increases her tips on average by about 1−3 dollars per dining party at Sunday brunch in this restaurant.

5. Solution: Activity 23-1 Marriage Ages

Pairing is effective here because of the variation in the ages of people who apply for marriage licenses. Look at the dotplots shown before part a: the ages extend from the teens to the seventies, with substantial overlap between the husband ages and wife ages. But there is a strong relationship between the ages of the husband and wife within a couple, as the graph in part c revealed. Thus, there is much less variation in the age *differences* than there is in the ages themselves. (It would be pretty surprising to find many couples with a 50-year age difference between the husband and the wife! But it would be much less surprising to find a husband and a wife from two different couples with such an age difference.) Note the small standard deviation of the differences (4.812 years) compared to the standard deviations of the ages (14.56 years for husbands and 13.56 years for wives). You first encountered this reduction in variability in Activity 9-6, and now you use this to your benefit—based on the principle you discovered in Activity 20-2 that a smaller standard deviation produces a more statistically significant result, when all else (in that case, the sample size and sample mean) remains the same. Researchers often think about increasing the sample size to decrease sampling variability, but here the researchers managed to decrease the variability in the measurements themselves with an effective study design.

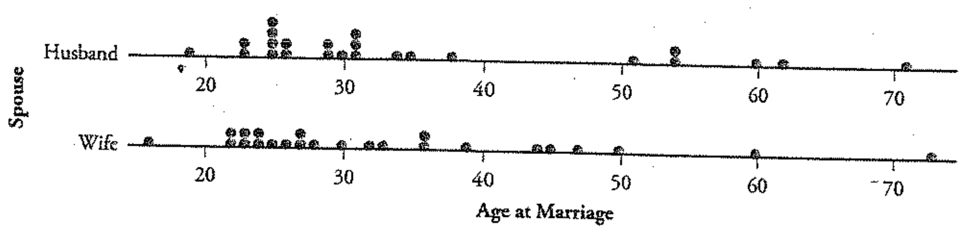
Watch Out

- Remember, the first analysis you conducted in this activity, using two-sample t-procedures, is completely wrong. (We asked you to perform the wrong procedure so you could see how much more powerful the correct procedure is.) It's crucial that you determine which t-procedure (paired or unpaired) is appropriate when you are comparing two groups on a quantitative response.
- Determining which t-procedure to use depends entirely on how the data are collected. If the sampling or experimental design is paired, then use paired t-procedures. But if the samples are drawn independently for the two groups, or if randomization is used to assign subjects to separate treatment groups, then use the two-sample t-procedures from Topic 22.
- To help decide whether the data are collected with a paired design, ask whether there's a link between each observation in one group with a specific observation in the other group. In this study, the link between the groups is each husband in one group is married to one specific wife in the other group. In a before/after study, the link between the groups is that the data are recorded on the same individuals.
- In a paired design, mixing up the order of values in one group would create a problem. With a nonpaired design, you can mix up the order of values in one group, or both groups, without affecting the analysis at all.
- If a study has a different number of observations between the two groups, then it can't be a paired design (unless data are not recorded for some observational units). However, even if the sample sizes are the same between the two groups, you cannot necessarily conclude the design is paired.

Reconsider the data presented in Activity 9-6 concerning the ages (in years) at marriage for a sample of 24 couples who obtained their marriage licenses in Cumberland County, Pennsylvania, in 1993. The data are recorded in the following table and stored in the data file `MarriageAges`. Also following are the graphical displays and summary statistics:

Couple #	Husband's Age	Wife's Age	Couple #	Husband's Age	Wife's Age	Couple #	Husband's Age	Wife's Age
1	25	22	9	31	30	17	26	27
2	25	32	10	54	[illegible]	18	31	[illegible]
3	51	50	11	23	25	19	26	[illegible]
4	25	25	12	34	39	20	62	60
5	38	33	13	25	24	21	29	26
6	30	27	14	23	22	22	31	23
7	60	45	15	19	16	23	29	28
8	54	47	16	71	73	24	35	36

	Sample Size	Sample Mean	Sample SD
Husband's Age	24	35.71	14.56
Wife's Age	24	33.83	13.56



a. Use these summary statistics to conduct a two-sample t-test of whether these sample data provide evidence that the population mean age of husbands exceeds that of wives. Report the hypotheses, test statistic, and p-value. Also state your test decision at the $\alpha = .05$ level and summarize your conclusion. (Feel free to use technology.)

$H_o: \mu_h = \mu_w$
$H_a: \mu_h > \mu_w$

$H_0: \mu_H = \mu_w$
$H_a: \mu_H > \mu_w$

$$t = \frac{\bar{x}_h - \bar{x}_w}{\sqrt{\frac{s_h^2}{n_h} + \frac{s_w^2}{n_w}}} \qquad t = \frac{35.71 - 33.83}{\sqrt{\frac{14.56^2}{24} + \frac{13.56^2}{24}}} = \frac{1.88}{4.0613} = .4629$$

P-value = .323

Fail to reject H_o. There is no evidence that on average the ages of husbands is greater than the ages of wives. There is a 32% chance of getting this difference in ages if we assume no difference.

b. Use a two-sample t-interval to estimate the difference in population mean ages with 90% confidence. (Feel free to use technology.)

$\bar{x}_H - \bar{x}_W \pm t^* \sqrt{\frac{s_H^2}{n_H} + \frac{s_W^2}{n_W}} = 35.71 - 33.83 \pm 1.714\left(\sqrt{\frac{14.56^2}{24} + \frac{13.56^2}{24}}\right)$

$1.88 \pm 1.714(4.0613)$

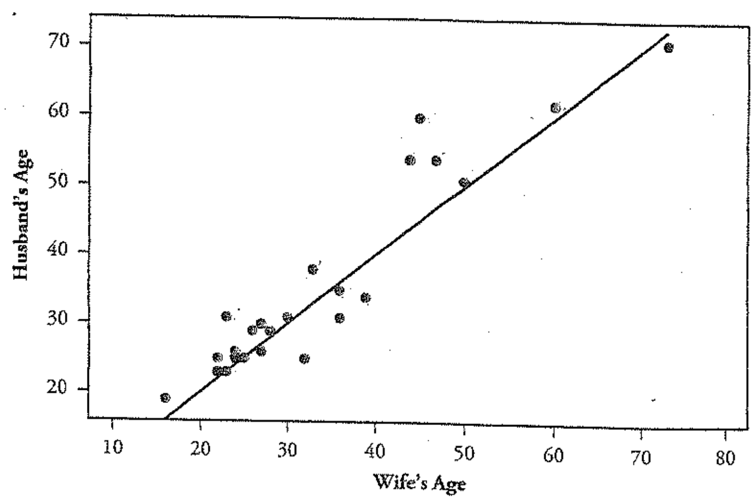
1.88 ± 6.9611

$\boxed{-5.0811 \text{ to } 8.8411}$

$\boxed{\text{calc}}$ -4.938 to 8.6983

Unfortunately, this analysis is completely inappropriate for these data. The two-sample t-procedure is not valid because these samples are not independent. This analysis would have been appropriate only if the researcher had gathered ages for one sample of 24 husbands and then independently gathered ages for a completely different sample of 24 wives.

c. Consider the following graph that displays the husband's age and wife's age for each couple. Notice that the graph includes a "$y = x$" line showing where the husband and wife have the same age. Comment on whether most of the values are above or below this line, and indicate whether there is a clear tendency for husbands to be older than their wives.



Slight tendency for husbands to be older than their wives.

12 clearly above

6 below

7 on line

d. Does this graph reveal a random scatter, or do the husband's and wife's ages appear to be related? In particular, do older people tend to marry older people and younger people tend to marry younger people? Explain.

Appears that older people marry older people + younger people marry younger people.

The key to devising a correct analysis of these data is realizing these data are *paired* because the observational units are *couples*, not independent individuals. The appropriate analysis is, therefore, to calculate the *differences* in ages for each couple, and then apply *one*-sample *t*-procedures (from Topics 19 and 20) to those differences.

A paired *t*-procedure applies one-sample *t*-procedures to the *differences* within a pair. The null hypothesis is $H_0: \mu_d = 0$, and the test statistic is

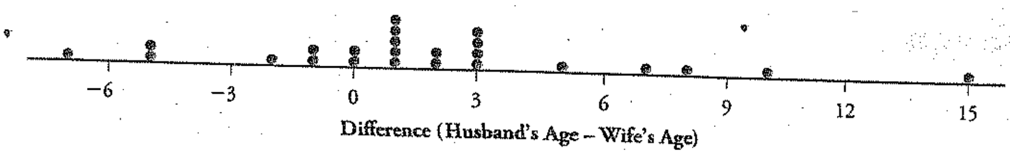
$$t = \frac{\bar{x}_d}{s_d/\sqrt{n}}$$

with a *p*-value based on the *t*-distribution with $(n - 1)$ degrees of freedom, where *n* is the number of *pairs* in the sample. (The subscript "*d*" reminds you that you are now analyzing *differences*.) A confidence interval for the population mean difference μ_d is

$$\bar{x}_d \pm t^* \frac{s_d}{\sqrt{n}}$$

The technical conditions are the same as with a one-sample *t*-procedure, except the observational units are pairs and the data are differences (e.g., a large sample size of pairs or the population of differences follows a normal distribution).

A dotplot and summary statistics for the age differences follow:



	Sample Size	Sample Mean	Sample SD
Difference	24	1.875	4.812

e. Determine a 90% confidence interval for the population mean *difference* in ages between husbands and wives. Also interpret this interval.

$\bar{x} \pm t^* \, s/\sqrt{n}$ $\quad 1.875 \pm 1.714 \, \frac{4.812}{\sqrt{24}}$

1.875 ± 1.6836

$\boxed{.1914 \text{ to } 3.5586}$ calc $\boxed{.19156 \text{ to } 3.5584}$

f. How does this interval compare to the (incorrect) interval from part b? Comment on its midpoint and width.

The midpoint, $\bar{x}$, are the same. The width is much narrower for the interval of differences.

g. Conduct a paired t-test of whether the sample data provide strong evidence that the population mean difference exceeds zero. Report the hypotheses, test statistic, and p-value. Also state your test decision at the $\alpha = .05$ level, and summarize your conclusion. (Feel free to use technology.)

$H_0: \mu = 0 \quad H_a: \mu > 0$

$\mu = \mu_{husband} - \mu_{wife}$

$$t = \frac{\bar{x} - 0}{s/\sqrt{n}} = \frac{1.875}{\frac{4.812}{\sqrt{24}}} = 1.909 \longrightarrow \boxed{.05 > p > .025}$$

P-value

$\boxed{\text{calc - P-value} = .0344}$

Reject Ho. There is evidence that husbands on average are older than their wives.

There is only a 3% chance of getting a difference in ages of 1.875 if we assume no difference.

h. How does your conclusion from this test compare to the (incorrect) conclusion from the test in part a?

Using the paired test - shows that husbands, on average, are older than their wives.

The other test does not show this.

i. Check that the technical conditions for applying these t-procedures are met.

SRS

normally distributed. - see dot plot of differences previous page.

j. Explain why the paired analysis produces such a different conclusion from the independent-samples analysis.

The paired test shows a significant difference in age because the st. dev. is much smaller when you look at the difference between a husband wife pair than the st. dev. between ages of all husbands & all wives.

k. Was the researcher wise for gathering paired data rather than independent-samples data to investigate this research question? In other words, was pairing helpful for estimating the population mean age difference among married couples? Explain.

Yes. see above.