

Course Notes for Statistics 301, Professor Wardrop

Chapter 2: Two Sampling distributions

The effect of sample size.

The sampling distribution of the test statistic for Fisher's test for the Chronic Crohn's Disease study:

x	$P(X = x)$	$P(X \leq x)$	$P(X \geq x)$
-0.46	0.0001	0.0001	1.0000
-0.41	0.0005	0.0006	0.9999
-0.35	0.0025	0.0031	0.9994
-0.29	0.0092	0.0123	0.9969
-0.24	0.0265	0.0388	0.9877
-0.18	0.0605	0.0993	0.9612
-0.12	0.1102	0.2095	0.9007
-0.07	0.1605	0.3700	0.7905
-0.01	0.1872	0.5572	0.6300
0.05	0.1752	0.7323	0.4428
0.10	0.1314	0.8637	0.2677
0.16	0.0788	0.9425	0.1363
0.21	0.0377	0.9802	0.0575
0.27	0.0143	0.9945	0.0198
0.33	0.0043	0.9988	0.0055
0.38	0.0010	0.9998	0.0012
0.44	0.0002	1.0000	0.0002

The sampling distribution of the test statistic for Fisher's test for the Ballerina study:

x	$P(X = x)$	$P(X \leq x)$	$P(X \geq x)$
-0.40	0.0009	0.0010	1.0000
-0.32	0.0081	0.0090	0.9990
-0.24	0.0387	0.0477	0.9910
-0.16	0.1127	0.1604	0.9523
-0.08	0.2104	0.3708	0.8396
0.00	0.2584	0.6292	0.6292
0.08	0.2104	0.8396	0.3708
0.16	0.1127	0.9523	0.1604
0.24	0.0387	0.9910	0.0477
0.32	0.0081	0.9990	0.0090
0.40	0.0009	1.0000	0.0010

Consider again the Infidelity Study. Many people find it surprising that such a large difference in \hat{p} 's is not statistically significant. As we will explore in this section, the large P-value is b/c of the small amount of data.

To keep this simple and yet to convey the essential results, I will restrict attention to studies for which the alternative is $p_1 > p_2$ and the observed value of the test statistic is positive ($x > 0$).

In the Infidelity Study, for the alternative $>$, the P-value is 0.1849.

The actual Infidelity Study, of course, had $n = 20$ subjects. Define the *imaginary doubled study* version of the Infidelity Study to have twice as many subjects (40) with all values in the table doubled. This would yield the following table of data.

	Tell?			
Cheater	Yes	No	Total	\hat{p}
Husband	14	6	20	0.70
Wife	8	12	20	0.40
Total	22	18	40	

Using the web site, the P-value for the alternative $>$ is 0.0555, which is smaller than the P-value for the actual study. Similarly, we can define the *imaginary tripled study* version of the Infidelity Study to have thrice as many subjects (60) with all values in the table tripled. This would yield the following table of data.

	Tell?			
Cheater	Yes	No	Total	\hat{p}
Husband	21	9	30	0.70
Wife	12	18	30	0.40
Total	33	27	60	

Using the web site, the P-value for the alternative $>$ is 0.0185, which is smaller than both the P-value for the actual study and the P-value for the imaginary doubled study.

Below is the general result.

Suppose that for the actual study, the alternative is $>$ and $x > 0$. Define the *imaginary k-times study* to yield the table one gets by multiplying each entry in the actual contingency table by k , where k is an integer larger than one.

Result: The larger the value of k , the smaller the P-value.

Note that this result is illustrated above for $k = 2$ and $k = 3$.

Example. Match each table with its P-value for the alternative $>$. The three P-values are 0.2095, 0.0622 and 0.3521.

Table A

Treat.	<i>S</i>	<i>F</i>	Total
1	6	4	10
2	10	12	22
Total	16	16	32

Table B

Treat.	<i>S</i>	<i>F</i>	Total
1	12	8	20
2	20	24	44
Total	32	32	64

Table C

Treat.	<i>S</i>	<i>F</i>	Total
1	30	20	50
2	50	60	110
Total	80	80	160

Solution. In order to apply our result, first find the table with the smallest n , in this case Table A. Note that for this table, $x > 0$. If we take A as the actual data, then Tables B and C are, respectively, the imaginary tables for $k = 2$ and $k = 5$. Thus, the P-value for A is 0.3521, for B is 0.2095, and for C is 0.0622.

Chapter 3: Simulation

Results of a simulation experiment with 10,000 runs for the Ballerina study:

<i>x</i>	Rel. Freq. of <i>x</i>	Rel. Freq. of $\leq x$	Rel. Freq. of $\geq x$
-0.40	0.0009	0.0009	1.0000
-0.32	0.0072	0.0081	0.9991
-0.24	0.0383	0.0464	0.9919
-0.16	0.1137	0.1601	0.9536
-0.08	0.2169	0.3770	0.8399
0.00	0.2591	0.6361	0.6230
0.08	0.2022	0.8383	0.3639
0.16	0.1140	0.9523	0.1617
0.24	0.0383	0.9906	0.0477
0.32	0.0089	0.9995	0.0094
0.40	0.0005	1.0000	0.0005

Results of a simulation experiment with 10,000 runs for the Crohn's study:

<i>x</i>	Rel. Freq. of <i>x</i>	Rel. Freq. of $\leq x$	Rel. Freq. of $\geq x$
-0.46	0.0002	0.0002	1.0000
-0.41	0.0005	0.0007	0.9998
-0.35	0.0027	0.0034	0.9993
-0.29	0.0094	0.0128	0.9966
-0.24	0.0289	0.0417	0.9872
-0.18	0.0593	0.1010	0.9583
-0.12	0.1178	0.2188	0.8990
-0.07	0.1540	0.3728	0.7812
-0.01	0.1893	0.5621	0.6272
0.05	0.1724	0.7345	0.4379
0.10	0.1287	0.8632	0.2655
0.16	0.0830	0.9462	0.1368
0.21	0.0345	0.9807	0.0538
0.27	0.0143	0.9950	0.0193
0.33	0.0039	0.9989	0.0050
0.38	0.0010	0.9999	0.0011
0.44	0.0001	1.0000	0.0001

Standard Normal Curve Approximation

Recall that $x = \hat{p}_1 - \hat{p}_2$ and

$$\sigma = \sqrt{\frac{m_1 m_2}{n_1 n_2 (n - 1)}}.$$

Without the continuity correction.

1. Compute $z = x/\sigma$.
2. For the
 - First alternative, $>$, the approximate P-value equals the area under the standard normal curve to the right of z .
 - Second alternative, $<$, the approximate P-value equals the area under the standard normal curve to the right of $-z$.
 - Third alternative, \neq , the approximate P-value equals **twice** the area under the standard normal curve to the right of $|z|$.

With the continuity correction.

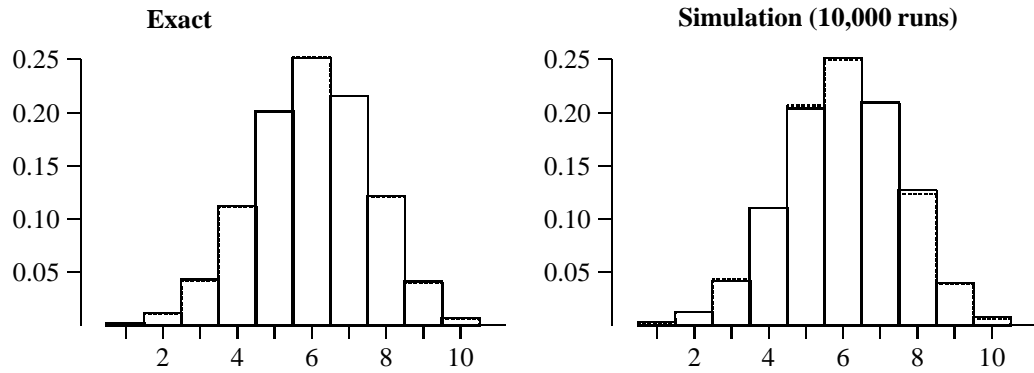
Compute:

$$g = \delta/2 = \frac{n}{2n_1 n_2}.$$

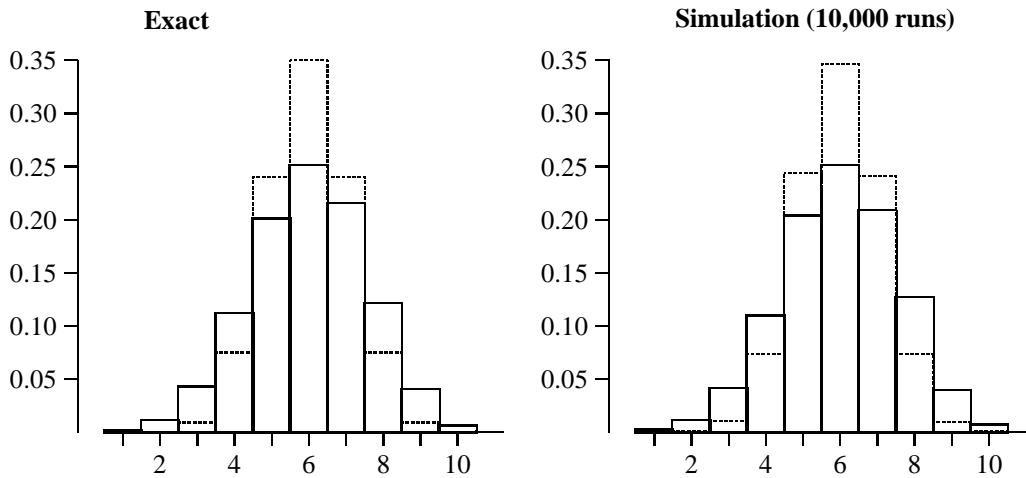
- For the first alternative, $>$,
 1. Compute $x_1 = x - g$ and $z_1 = x_1/\sigma$.
 2. The approximate P-value equals the area under the standard normal curve to the right of z_1 .
- For the second alternative, $<$,
 1. Compute $x_2 = x + g$ and $z_2 = x_2/\sigma$.
 2. The approximate P-value equals the area under the standard normal curve to the right of $-z_2$.
- For the third alternative, \neq , if $|x| \leq g$, then the exact P-value equals one and no approximation is needed; otherwise,
 1. Compute $x_3 = |x| - g$ and $z_3 = x_3/\sigma$. (Note that both x_3 and z_3 are larger than zero.)
 2. The approximate P-value equals **twice** the area under the standard normal curve to the right of z_3 .

Chapter 5: Sampling With or Without Replacement

Two probability histograms for X , the number of successes in a sample of size $n = 10$ from a dichotomous box with $N = 1,000$ and $p = 0.6$. Solid [Dashed] rectangles are for a random sample with [without] replacement.



Two probability histograms for X , the number of successes in a sample of size $n = 10$ from a dichotomous box with $N = 20$ and $p = 0.6$. Solid [Dashed] rectangles are for a random sample with [without] replacement.



Chapter 7: Background (lurking) variables. A company with 200 employees decides it must reduce its work force by one-half. The following table reveals the relationship between gender and outcome.

Gender	Outcome		Total	\hat{p}
	Released	Not released		
Female	60	40	100	0.60
Male	40	60	100	0.40
Total	100	100	200	

In an observational study, we should consider the possible influence of a background variable. For illustration, suppose that the company has workers in two job classifications, A and B. Consider four scenarios.

Case 1					Case 1				
Job A					Job B				
Gender	Outcome		Total	\hat{p}	Gender	Outcome		Total	\hat{p}
	Released	Not released				Released	Not released		
Female	30	20	50	0.60	Female	30	20	50	0.60
Male	20	30	50	0.40	Male	20	30	50	0.40
Total	50	50	100		Total	50	50	100	

Case 2					Case 2				
Job A					Job B				
Gender	Outcome		Total	\hat{p}	Gender	Outcome		Total	\hat{p}
	Released	Not released				Released	Not released		
Female	30	10	40	0.75	Female	30	30	60	0.50
Male	30	30	60	0.50	Male	10	30	40	0.25
Total	60	40	100		Total	40	60	100	

Case 3					Case 3				
Job A					Job B				
Gender	Outcome		Total	\hat{p}	Gender	Outcome		Total	\hat{p}
	Released	Not released				Released	Not released		
Female	60	15	75	0.80	Female	0	25	25	0.00
Male	40	10	50	0.80	Male	0	50	50	0.00
Total	100	25	125		Total	0	75	75	

Case 4					Case 4				
Job A					Job B				
Gender	Outcome		Total	\hat{p}	Gender	Outcome		Total	\hat{p}
	Released	Not released				Released	Not released		
Female	56	24	80	0.70	Female	4	16	20	0.20
Male	16	4	20	0.80	Male	24	56	80	0.30
Total	72	28	100		Total	28	72	100	

Chapter 12:

Traffic Data

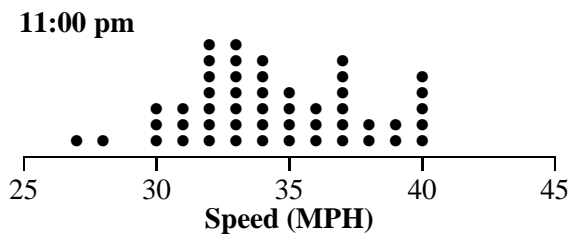
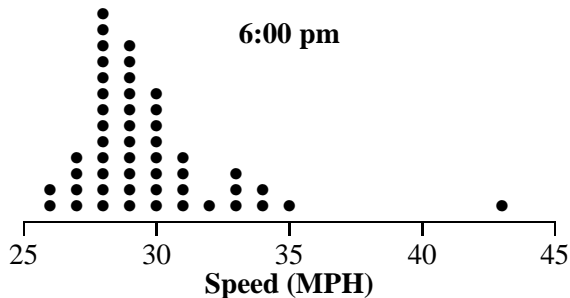
On a spring evening, a Milwaukee police officer measured the speeds of 100 automobiles. The data were collected on a street in a “warehouse district” with a speed limit of 25 MPH. Fifty cars were measured between roughly 5:45 and 6:15 pm, referred to below as 6:00 pm. The remaining 50 cars were measured between roughly 10:40 and 11:20 pm, referred to below as 11:00 pm.

Each car’s speed was measured to the nearest MPH. The sorted data, by time, are below.

Speeds at 6:00 pm									
26	26	27	27	27	27	28	28	28	28
28	28	28	28	28	28	28	28	28	29
29	29	29	29	29	29	29	29	29	29
30	30	30	30	30	30	30	30	31	31
31	31	32	33	33	33	34	34	35	43

Speeds at 11:00 pm									
27	28	30	30	30	31	31	31	32	32
32	32	32	32	32	33	33	33	33	33
33	33	34	34	34	34	34	34	35	35
35	35	36	36	36	37	37	37	37	37
37	38	38	39	39	40	40	40	40	40

Below are dot plots of these data.



The sum of the 50 speeds at 6:00 pm is 1484; thus, the mean is $1484/50 = 29.68$ MPH. The sum of the 50 speeds at 11:00 pm is 1721; thus, the mean is $1721/50 = 34.42$ MPH. The mean is $34.42 - 29.68 = 4.74$ MPH higher at the later time.

The sample sizes, both 50, are even; thus, there are two middle positions, position numbers 25 and 26. For 6:00 pm, the numbers in these positions are both 29; hence, the median is 29 MPH. For 11:00 pm, the numbers in these positions are both 34; hence, the median is 34 MPH. The median is $34 - 29 = 5$ MPH higher at the later time.

The range at 6:00 pm is $43 - 26 = 17$ MPH; at 11:00 pm it is $40 - 27 = 13$ MPH. Thus, according to the range, the data collected at 6:00 pm have more spread than the data collected at 11:00 pm.

The quartiles for 6:00 pm are 28 and 30, giving an IQR of $30 - 28 = 2$ MPH. For 11:00 pm, they are 32 and 37, giving an IQR of $37 - 32 = 5$ MPH. Thus, according to the IQR the data collected at 6:00 pm have less spread than the data collected at 11:00 pm. Thus, these two measures disagree on the relative spread in the two data sets.

The standard deviation for the 6:00 pm data is 2.81 MPH and for the 11:00 pm data is 3.25. For 6:00 pm, the one sd interval is

$$29.68 - 2.81 = 26.87 \text{ to } 29.68 + 2.81 = 32.49.$$

By counting, 41 of 50 observations, or 82%, are in the interval. This is considerably larger than the 68% predicted by the empirical rule. Why? The large outlier inflates the standard deviation. For example, if the outlier is deleted, the mean becomes 29.41, the standard deviation becomes 2.07 and the one sd interval becomes $[27.34, 31.48]$. Thirty-six of 49 (73.5%) observations are in this interval.

For 11:00 pm, the one sd interval is $[31.17, 37.67]$; 33 of 50 (66%) observations are in this interval.

Chapter 12: Percentiles and Quantiles

The material of this section is NOT in the text and will NOT be covered in lecture, but it is needed for some homework and practice exercises and it might be on the final exam. You may ask about this material at discussion, at the review for the final or during office hours.

The mean and standard deviation are arithmetic summaries of data; i.e. we get them by performing arithmetic operations: adding, dividing, and so on. Percentiles and quantiles are not arithmetic summaries; rather, they focus on positions in the sorted list of data.

The median is also the 50th percentile and the 0.50 quantile. By convention, there are 99 percentiles, namely, the 1st, 2nd, 3rd, ..., and 99th. Quantiles are essentially the decimal version of a percentile. For example, the 63rd percentile is the same as the 0.63 quantile.

To make this abstract, but I hope not confusing, let π be a number strictly between 0 and 1; i.e. $0 < \pi < 1$. The π quantile is the same number as the 100π percentile. My example of the previous paragraph illustrates this notion with $\pi = 0.63$; the 0.63 quantile is the same number as the $100(0.63) = 63$ rd percentile.

By convention, there are 99 percentiles, allowing π to be any of the values: 0.01, 0.02, 0.03, ... 0.99.

As stated earlier, the median is the 0.50 quantile and the 50th percentile. We will explore this idea further now. Consider a sample of three distinct numbers sorted, for example: 4, 9 and 20. The median is 'the number in the center position,' which is 9. For a sample of four distinct numbers sorted, for example: 4, 7, 9 and 20. There are now two center positions, which contain the numbers 7 and 9. By convention, the median is taken to be the arithmetic average of these numbers, 8, which has the property that half the data are smaller than it and half are larger than it. So what is the median? Well, there is no one simple answer. But there is one not-so-simple answer, which we have not seen yet.

It turns out that for math purposes, the idea 'half above, half below' is closer to what we want than the idea 'in the center.' Thus, we define the median as

follows.

The median is any number with the following two properties.

- *At least* one-half of the data are *less than or equal to* the median.
- *At least* one-half of the data are *greater than or equal to* the median.

For any odd sample size, for example our data 4, 9 and 20, there is unique number that satisfies this definition, namely the 9. Let's check it.

- Two-thirds of the data are less than or equal to 9.
- Two-thirds of the data are greater than or equal to 9.

Thus, 9 satisfies our definition. But is it unique? Yes. (For any candidate smaller than 9, the first item fails; for any candidate larger than 9, the second item fails.)

Next, consider an even sample size, for example our earlier data of 4, 7, 9 and 20. You can check that our median 8 satisfies both conditions. But it is not unique. Actually, any number between 7 and 9 inclusive will satisfy our definition of the median. Also, any number smaller than 7 or larger than 9 will fail our definition. Thus, in a strictly literal math sense, for an even sample size there can be an entire interval of numbers that satisfy the definition of median. But in order to avoid being labeled as a group of needlessly sadistic people, except perhaps when proving theorems, all statisticians and mathematicians agree that if there is an interval of medians, we call the midpoint of the interval *the median*. And we will always do this in this course.

You might well wonder what was the point of all of the above. Before you read this we all agreed on what the median was, and now I have just made it more complicated. Well, we need the above for percentiles and quantiles. I will illustrate.

First, let us be specific. Suppose we want to find the 35th percentile, which we will denote by $P_{35} = Q_{0.35}$, the 0.35 quantile. We define the 35th percentile to be any number that has the following two properties.

- At least 35% of the data are *less than or equal to* it.
- At least 65% of the data are *greater than or equal to* it.

If an entire interval of numbers satisfy both properties, we take the 35th percentile to be the midpoint of the interval.

Next, I give you the algorithm for calculating the 35th percentile.

1. Calculate $k = 0.35n$.
2. If k is an integer, then the 35th percentile equals

$$[x_{(k)} + x_{(k+1)}]/2.$$

3. If k is not an integer, round it up to the next integer and call it k' . The 35th percentile equals $x_{(k')}$.

For example, suppose $n = 100$. Then,

$$0.35n = 0.35(100) = 35$$

is an integer and the 35th percentile equals

$$[x_{(35)} + x_{(36)}]/2.$$

Let's check that this works.

First, suppose that $x_{(35)}$ and $x_{(36)}$ are different numbers. Then exactly 35% of the data are less than the percentile and exactly 65% of the data are greater than the percentile. If $x_{(35)}$ and $x_{(36)}$ are the same number then at least 35% of the data are less than or equal to the percentile and at least 65% of the data are greater than or equal to the percentile.

As another example, suppose that $n = 150$. Then

$$k = 0.35n = 0.35(150) = 52.5$$

is not an integer, so we round it up to $k' = 53$ and the 35th percentile equals $x_{(53)}$. Let's check that it works.

First, clearly at least 53 observations are less than or equal to $x_{(53)}$, and $53/150 = 0.353$ is at least 35%. Second, at least $150 - 52 = 98$ observations are

greater than or equal to $x_{(53)}$, and $98/150 = 0.653$ is at least 65%

Now that we understand the 35th percentile, we will consider any arbitrary percentile. First, the definition. We define the 100π percentile to be any number that has the following two properties.

- At least $100\pi\%$ of the data are *less than or equal to* it.
- At least $100(1-\pi)\%$ of the data are *greater than or equal to* it.

If an entire interval of numbers satisfy both properties, we take the 100π percentile to be the midpoint of the interval.

Next, I give you the algorithm for calculating the 100π percentile.

1. Calculate $k = 100\pi n$.
2. If k is an integer, then the 100π percentile equals

$$[x_{(k)} + x_{(k+1)}]/2.$$

3. If k is not an integer, round it up to the next integer and call it k' . The 100π percentile equals $x_{(k')}$.

Final comment. The 25th percentile is called the first quartile and the 75th percentile is called the third quartile. The text presents a method different than the one in this section for computing quartiles. Actually, the two methods give the same answers *unless* the sample size n is equal to $4m + 1$, where m is a positive integer. In other words, the two methods give the same answer unless

$$n = 5, 9, 13, \dots$$

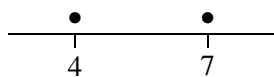
For homework and project 3, you may use either method for calculating quartiles. The final will *not* ask you to calculate quartiles.

Chapter 12: The Standard Deviation

This material on the standard deviation is optional and is not required for the exams or future lectures.

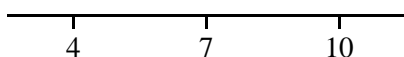
It is provided for the student who is interested in this topic.

Consider any two numbers, for example 4 and 7. Think of subtraction as a way to compare two numbers. First, $7 - 4 = 3$, which means that 7 is 3 units larger than 4. Second, $4 - 7 = -3$, which means that 4 is 3 units smaller than 7. Next, think of these numbers on a dot plot, as below.



The equation $7 - 4 = 3$ means that 7 is 3 units to the right of 4 (a positive difference means to the right). Similarly, $4 - 7 = -3$ means that 4 is 3 units to the left of 7 (a negative difference means to the left). The absolute value of $7 - 4$ is written as $|7 - 4|$ and equals 3. The absolute value is the distance between the two numbers. (Remember that distance cannot be negative, and is zero if, and only if, we compare a number to itself; for example, $|5 - 5| = 0$.)

Consider the interval of numbers from 4 to 10, inclusive, written $[4, 10]$.



The center, or midpoint, of this interval is at 7. Convince yourself of the truth of the following fact: If b is a number in the interval $[4, 10]$, then the distance between b and 7 is 3 units or less. Another way to say this is: Every number in the interval $[4, 10]$ is within 3 units of 7. In general, if $[c, d]$ is an interval of numbers with midpoint equal to m ($m = (c + d)/2$), width w ($w = d - c$) and half-width h ($h = w/2$), then every number in the interval is within h units of the midpoint m .

Exercise 5 on page 403 of the text describes a study, by Kymn, that yielded the following sorted data: 489, 490, 492, 493, 493 (seconds). I will use these data to illustrate the computation of the variance, denoted by s^2 , and its square root, s , the standard deviation (sd).

First, we compute the mean of the data:

$$\frac{489 + 490 + 492 + 493 + 493}{5} = \frac{2457}{5} = 491.4$$

seconds. Next, we construct the following table.

x	$x - \bar{x}$	$(x - \bar{x})^2$
489	-2.4	5.76
490	-1.4	1.96
492	0.6	0.36
493	1.6	2.56
493	1.6	2.56
Total	0.0	13.20

Each $(x - \bar{x})$ is called a deviation and each $(x - \bar{x})^2$ is called a squared deviation.

The variance is

$$s^2 = 13.20/4 = 3.30$$

seconds squared, and the sd is

$$s = \sqrt{s^2} = \sqrt{3.30} = 1.817 \text{ seconds.}$$

The formula for the variance is

$$s^2 = \frac{\sum(x - \bar{x})^2}{n - 1}.$$

Call:

- the interval $[\bar{x} - s, \bar{x} + s]$ the one sd interval,
- the interval $[\bar{x} - 2s, \bar{x} + 2s]$ the two sd interval, and
- the interval $[\bar{x} - 3s, \bar{x} + 3s]$ the three sd interval.

Note that all data points in the k ($k = 1, 2$ or 3) sd interval are within k sd's of the mean. For Kymn's data, the one, two, and three sd intervals are:

- $[491.4 - 1.82, 491.4 + 1.82] = [489.58, 493.22]$,
- $[491.4 - 3.63, 491.4 + 3.63] = [487.77, 495.03]$, and
- $[491.4 - 5.45, 491.4 + 5.45] = [485.95, 496.85]$.

Notice that 4 of 5 observations are within one sd of the mean and all 5 are within two sd's of the mean.

Look again at the computation of the sd for Kymn's data. The data are integers, which make arithmetic pleasant, but the deviations have one digit after the decimal point and the squared deviations have two digits after the decimal point. We can avoid

these fractions by using an alternate formula for the variance:

$$s^2 = \frac{n \sum x^2 - (\sum x)^2}{n(n-1)}.$$

This formula is illustrated via the following table.

x	x^2
489	239121
490	240100
492	242064
493	243049
493	243049
Total	2457 1207383

Thus,

$$s^2 = \frac{5(1207383) - (2457)^2}{5(4)} =$$

$$\frac{6036915 - 6036849}{20} = \frac{66}{20} = 3.30,$$

as before.

This alternate formula is unsatisfactory for Kymn's data because the numbers, while integers, are very large. This difficulty is avoided by **recoding** the data. An easy way to recode is to subtract the smallest observation, in this case 489, from each observation. This changes the data set to:

0, 1, 3, 4, and 4.

The appropriate alternate table for these data is below.

x	x^2
0	0
1	1
3	9
4	16
4	16
Total	12 42

Thus,

$$s^2 = \frac{5(42) - (12)^2}{5(4)} = \frac{210 - 144}{20} = \frac{66}{20} = 3.30,$$

as before.

If there are numerous ties in the data set, the formula for the variance can be modified to save time. I will illustrate with the 6:00 traffic data. I will begin by subtracting 26 from each observation. The resulting data are below, where f is for frequency.

x	f	fx	x^2	fx^2
0	2	0	0	0
1	4	4	1	4
2	13	26	4	52
3	11	33	9	99
4	8	32	16	128
5	4	20	25	100
6	1	6	36	36
7	3	21	49	147
8	2	16	64	128
9	1	9	81	81
17	1	17	289	289
Total	50	184	—	1064

The variance is

$$s^2 = \frac{n(\sum fx^2) - (\sum fx)^2}{n(n-1)}.$$

For Kenny's 6:00 data,

$$s^2 = \frac{50(1064) - (184)^2}{50(49)} =$$

$$\frac{53200 - 33856}{2450} = \frac{19344}{2450} = 7.8955,$$

and the sd is $s = \sqrt{7.8955} = 2.81$ MPH.

If we want to drop the outlier from the data set, we simply delete the last row of the table above. With this change,

$$\sum fx = 167 \text{ and } \sum fx^2 = 775.$$

The variance becomes

$$s^2 = \frac{49(775) - (167)^2}{49(48)} =$$

$$\frac{37975 - 27889}{2352} = \frac{10086}{2352} = 4.2883,$$

and the sd is $s = \sqrt{4.2883} = 2.07$ MPH. Note that the deletion of one observation (out of 50) decreased the standard deviation by over 26% (from 2.81 to 2.07).

Chapter 15: The confidence interval for the population mean

This material is optional; see the above comment about the standard deviation. The reader might, however, want to read the examples below.

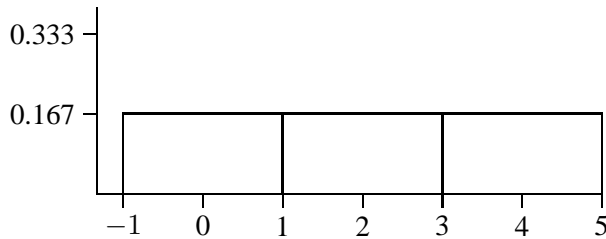
Recall that a population is a picture—a probability histogram for a count response, and a pdf for a measure response. In either case, the mean of the population is the center of gravity of the picture and is denoted by μ .

Suppose that the scientist's objective is to estimate the mean of the population. Assume that we have a random sample from the population. Let \bar{x} and s denote the mean and standard deviation of the data.

The point estimate of the population mean is the sample mean. Point estimates are usually wrong, but how good/bad are they? Is the point estimate *probably close* to μ ? Or is it *likely to be very different* from μ ? These questions are answered by examining the sampling distribution of the point estimate, \bar{x} .

Recall that the sampling distribution of \bar{x} is a listing of all possible values of \bar{x} and the probability of each of these values. Alternatively, this "listing" can be presented as a picture.

We will consider an extremely simple example. The population consists of three subjects, whose responses are 0, 2, and 4. Below is a picture of this population.



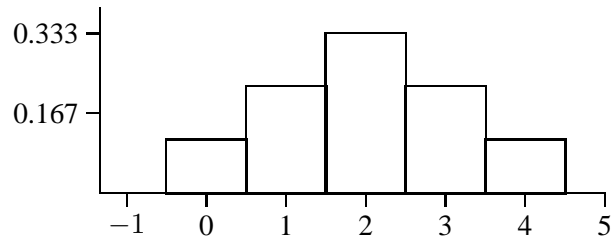
Clearly, the center of gravity of this picture is 2; thus, $\mu = 2$. It can be shown (see Section 3.6 of the text) that the standard deviation of the population is

$$\sigma = \sqrt{8/3} = 1.633.$$

Next, we will derive the sampling distribution for \bar{x} for a sample of size two. We begin by listing all possible samples of size two.

Sample	Values	\bar{x}	Sample	Values	\bar{x}
1	0, 0	0	6	2, 4	3
2	0, 2	1	7	4, 0	2
3	0, 4	2	8	4, 2	3
4	2, 0	1	9	4, 4	4
5	2, 2	2			

The probability histogram for this distribution is below.



Clearly, the center of gravity of this sampling distribution is 2. It can be shown that the standard deviation of this sampling distribution is $\sqrt{4/3} = 1.1547$. To summarize,

	Population	\bar{x} for $n = 2$
Center	$\mu = 2$	$\mu = 2$
Spread	$\sigma = 1.633$	$\sigma/\sqrt{2} = 1.1547$

This table reflects the following **mathematical facts** that are true for any (every) sample size n .

- The center of gravity of the sampling distribution of \bar{x} equals μ .
- The standard deviation (called standard error) of the sampling distribution of \bar{x} equals σ/\sqrt{n} .

The consequence of these facts is that it is easy to standardize \bar{x} ,

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}.$$

Finally, the pictures above illustrate the following *qualitative* mathematical fact.

- As n becomes larger, the sampling distribution of \bar{x} becomes more bell-shaped (and symmetric).

We now have the necessary ingredients to obtain a confidence interval for μ . Look at the expression for z ,

$$\frac{\bar{x} - \mu}{\sigma/\sqrt{n}},$$

and note its similarity to the standardized version of \hat{p} that was given in chapter 6:

$$\frac{\hat{p} - p}{\sqrt{pq/n}}$$

First, each expression is a ratio. In the numerator we have the estimate (either \bar{x} or \hat{p}) compared (via subtraction) to the quantity we are trying to estimate (either μ or p). Each denominator is a measure of spread (we do not need to be more precise) which contains a quantity whose value is unknown (σ or pq).

In Chapter 6 we learned that Slutsky's work showed that one could replace the unknown pq in the denominator by the computable $\hat{p}\hat{q}$, and the sampling distribution would remain relatively unchanged; that is, it could still be well-approximated by the standard normal curve. In other words, we learned that it is reasonable to approximate the sampling distribution of

$$(\hat{p} - p)/\sqrt{\hat{p}\hat{q}/n},$$

with the standard normal curve. This led to the formula

$$\hat{p} \pm z\sqrt{\hat{p}\hat{q}/n},$$

as the confidence interval for p .

Slutsky's work can also be applied to the problem of Chapter 15, but the result is somewhat messier, as discussed below. In particular, Slutsky showed that one can replace the unknown σ with the computable s and use the standard normal curve to approximate the sampling distribution of

$$\frac{\bar{x} - \mu}{s/\sqrt{n}}.$$

Using the algebra from Chapter 6 (with different symbols; \bar{x} now instead of \hat{p} then, and so on), we find that

$$\bar{x} \pm zs/\sqrt{n}$$

is Slutsky's confidence interval for μ .

Example: Wendy selects a random sample of size $n = 144$ from a population and computes $\bar{x} = 83.2$ and $s = 21.4$. Find Slutsky's 95 percent confidence interval for μ .

Solution: For 95 percent confidence, $z = 1.96$. Thus, the confidence interval is

$$83.2 \pm 1.96(21.4)/\sqrt{144} =$$

$$83.2 \pm 3.5 = [79.7, 86.7].$$

Example: Walt selects a random sample of size $n = 225$ from a population and computes $\bar{x} = 126.8$ and $s = 38.7$. Find Slutsky's 90 percent confidence interval for μ .

Solution: For 90 percent confidence, $z = 1.645$. Thus, the confidence interval is

$$126.8 \pm 1.645(38.7)/\sqrt{225} =$$

$$126.8 \pm 4.2 = [122.6, 131.0].$$

After mastering this computation, it is natural to wonder, "How good is the answer?" The answer is based on using the standard normal curve as an approximating device; thus, my real interest is in whether using the standard normal curve is a good idea. Statisticians have been concerned about this question for over 100 years; below is a summary of what we have learned.

If the sample size is small, then the above method does not work very well. The general consensus is that "small" means 30 or fewer. If the sample size is larger than 30, then it is *sometimes* the case that the above method works well and *sometimes* the case that it works poorly. Most writers of introductory statistics texts don't want to explain what "sometimes" means and instead tell the readers that the above method always works fine if n is larger than 30.

In any event, I will now describe the method people use if n is 30 or fewer.

Gosset studied this problem and his work appeared in 1908. To improve upon Slutsky's result, Gosset needed to make an additional assumption about his population. He assumed that the population was a bell-shaped curve. Let me contrast what Slutsky found with what Gosset found.

Both men were interested in the sampling distribution of

$$\frac{\bar{x} - \mu}{s/\sqrt{n}}.$$

Slutsky showed that we could approximate the sampling distribution by using the standard normal curve. Gosset, with his extra assumption, was able to *write down* the exact sampling distribution! There are two salient features of Gosset's work:

1. The sampling distribution varies with n . This means that the sampling distribution for $n = 10$, for example, is different than the sampling distribution for $n = 11$.
2. Gosset's sampling distributions are called the t -distributions with $(n - 1)$ degrees of freedom (df).

For example, for a random sample of size $n = 12$, then the reference distribution is the t -distribution with 11 df; if the sample size is $n = 16$, the reference distribution is the t -distribution with 15 df.

In Slutsky's confidence interval for μ ,

$$\bar{x} \pm zs/\sqrt{n},$$

z is a number that comes from the reference curve (the standard normal curve). Gosset's confidence interval is

$$\bar{x} \pm ts/\sqrt{n},$$

with t a number that comes from the reference curve, which in this case is the t -distribution with $(n - 1)$ df. The appropriate value of t can be found in Table A.6 in our text, which appears on page 655 and also on the inside of the back cover.

Example: Sally selects a random sample of size $n = 14$ from a population and computes $\bar{x} = 73.2$ and $s = 11.4$. Find Gosset's 95 percent confidence interval for μ .

Solution: For 95 percent confidence, $t = 2.160$, which is found in row 13 (df) and the 95% column (see the bottom row) of Table A.6. Thus, the confidence interval is

$$73.2 \pm 2.160(11.4)/\sqrt{14} =$$

$$73.2 \pm 6.6 = [66.6, 79.8].$$

Section 16.2: Population inference for two numerical populations

Recall from lecture that $(\bar{X} - \bar{Y})$ is the point estimate of $(\mu_X - \mu_Y)$, and that the standardized version of $(\bar{X} - \bar{Y})$ is

$$W = \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{n_1} + \frac{\sigma_Y^2}{n_2}}}.$$

There are three ways statisticians use W to compare the population means. These ways are called Cases 1, 2, and 3 in the text. You will be responsible for Cases 1 and 3 only.

In Case 1, we assume that the two populations have the same variance, $\sigma_X^2 = \sigma_Y^2$. We call this common value of the variance σ^2 . We estimate σ^2 by

$$s_p^2 = \frac{(n_1 - 1)s_X^2 + (n_2 - 1)s_Y^2}{n_1 + n_2 - 2}.$$

The confidence interval for $(\mu_X - \mu_Y)$ is below, where the df for t is $n_1 + n_2 - 2$.

$$(\bar{x} - \bar{y}) \pm ts_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}.$$

To test $H_0 : \mu_X = \mu_Y$ the test statistic is:

$$t_1 = \frac{\bar{x} - \bar{y}}{s_p \sqrt{(1/n_1) + (1/n_2)}}.$$

(The subscript of 1 reminds us that we are using the first case.)

Example 1: Biking study. The data in this example come from a class project submitted by Sheryl.

A trial consisted of Sheryl performing a 1.5 mile sprint on her bicycle. In treatment 1, Sheryl loaded her pannier with 20 pounds and in treatment 2 she removed her pannier from her bike. The response is the time, in seconds, Sheryl required to complete the sprint.

In order to analyze her data, we will assume that we have random samples from two populations. Sheryl's data yielded the following summary statistics:

$$\bar{x} = 383.2, s_X = 4.38, n_1 = 5, \bar{y} = 355.2,$$

$$s_Y = 4.87, \text{ and } n_2 = 5.$$

We begin our analysis by computing s_p^2 .

$$s_p^2 = \frac{4(4.38)^2 + 4(4.87)^2}{5 + 5 - 2} =$$

$$\frac{4(19.18) + 4(23.72)}{8} = 21.45.$$

Thus, $s_p = \sqrt{21.45} = 4.63$.

The 95% confidence interval for $(\mu_X - \mu_Y)$ is

$$(383.2 - 355.2) \pm 2.306(4.63)\sqrt{1/5 + 1/5} = \\ 28.0 \pm 6.75 = [21.25, 34.75].$$

In words, based on my confidence interval, I conclude that Sheryl's mean time increases by between 21.25 and 34.75 seconds when the weighted pannier is added to her bike.

Next, I test the null hypothesis against the alternative that $\mu_X > \mu_Y$. (I consider it inconceivable that adding weight could reduce the mean.) The observed value of the test statistic is:

$$t_1 = \frac{28.0}{4.63\sqrt{1/5 + 1/5}} = 9.56.$$

This value, 9.56, is much greater than the largest entry in row $df = 8$, which is 3.355. Thus, the P-value is smaller than 0.005. (In fact, the area to the right of 9.56 under the t-curve with 8 df is 0.00000006, 6 in 100 million.)

Example 2: Cat treat study. The data in this example come from a class project performed by Dawn.

A trial consisted of Dawn placing 10 cat treats in front of her cat Bob. In treatment 1, the treats were chicken-flavored and in treatment 2 they were tuna-flavored. The response is the number of treats Bob eats in 10 minutes.

In order to analyze her data, we will assume that we have random samples from two populations. Dawn's data yielded the following summary statistics:

$$\bar{x} = 5.1, s_X = 2.025, n_1 = 10, \bar{y} = 2.9, \\ s_Y = 2.079, \text{ and } n_2 = 10.$$

We begin our analysis by computing s_p^2 .

$$s_p^2 = \frac{9(2.025)^2 + 9(2.079)^2}{10 + 10 - 2} = \\ \frac{9(4.10) + 9(4.32)}{18} = 4.21.$$

Thus, $s_p = \sqrt{4.21} = 2.052$.

The 95% confidence interval for $(\mu_X - \mu_Y)$ is

$$(5.1 - 2.9) \pm 2.101(2.052)\sqrt{1/10 + 1/10} = \\ 2.2 \pm 1.93 = [0.27, 4.13].$$

In words, based on my confidence interval, I conclude that Bob's mean consumption of treats increases by between 0.27 and 4.13 when he is offered chicken rather than tuna flavor.

Next, I test the null hypothesis against the alternative that $\mu_X \neq \mu_Y$. (I consider both possibilities to be conceivable.) The observed value of the test statistic is:

$$t_1 = \frac{2.2}{2.025\sqrt{1/10 + 1/10}} = 2.43.$$

This area to the right of 2.43 under the t-curve with 18 df is between 0.01 and 0.025. Thus, the P-value for the third alternative is between 0.02 and 0.05.

Case 3. In Case 3, we make no assumption about the values of the two population variances. Thus, we do not compute s_p^2 in Case 3. It is my advice to use Case 3 only if both sample sizes are 20 or more.

The Case 3 confidence interval for $(\mu_X - \mu_Y)$ is:

$$(\bar{x} - \bar{y}) \pm z\sqrt{\frac{s_X^2}{n_1} + \frac{s_Y^2}{n_2}}.$$

To test $H_0 : \mu_X = \mu_Y$ the test statistic is:

$$z = (\bar{x} - \bar{y}) / \sqrt{\frac{s_X^2}{n_1} + \frac{s_Y^2}{n_2}}.$$

Example 3: Batting study. The data in this example come from a class project submitted by Luke.

A trial consisted of Luke hitting a baseball. In treatment 1, he used an aluminum bat and in treatment 2 he used a wooden bat. The response is the distance, in feet, that Luke hit the ball.

In order to analyze his data, we will assume that we have random samples from two populations. Luke's data yielded the following summary statistics:

$$\bar{x} = 179.6, s_X = 62.1, n_1 = 40, \bar{y} = 166.2,$$

$$s_Y = 54.2, \text{ and } n_2 = 40.$$

The 95% confidence interval for $(\mu_X - \mu_Y)$ is

$$(179.6 - 166.2) \pm 1.96 \sqrt{\frac{(62.1)^2}{40} + \frac{(54.2)^2}{40}} =$$

$$13.4 \pm 25.5 = [-12.1, 38.9].$$

In words, based on my confidence interval, Luke's data are inconclusive. The mean with the aluminum bat is between 12.1 feet smaller and 38.9 feet larger than the mean with the wooden bat.

Next, I test the null hypothesis against the alternative that $\mu_X > \mu_Y$. (I consider it inconceivable that wood would have a larger mean than aluminum.) The observed value of the test statistic is:

$$z = 13.4/13.03 = 1.03.$$

The P-value is 0.1515.

Example 4: Math scores study. The data in this example come from a class project submitted by LDR (Lindsay, Dan and Rebecca).

A trial consisted of a student taking a math exam. The students came from two sources, an advanced class (1) and a remedial class (2).

In order to analyze these data, we will assume that we have random samples from two populations. LDR's data yielded the following summary statistics:

$$\bar{x} = 4.39s_X = 1.761, n_1 = 36, \bar{y} = 3.72,$$

$$s_Y = 1.861, \text{ and } n_2 = 36.$$

The 95% confidence interval for $(\mu_X - \mu_Y)$ is

$$(4.39 - 3.72) \pm 1.96 \sqrt{\frac{(1.761)^2}{36} + \frac{(1.861)^2}{36}} =$$

$$0.67 \pm 0.84 = [-0.17, 1.51].$$

In words, based on my confidence interval, LDR's data are inconclusive. The mean for the advanced class is between 0.17 points smaller and 1.51 points larger than the mean for the remedial.

Next, I test the null hypothesis against the alternative that $\mu_X > \mu_Y$. (I consider it inconceivable

that the remedial class would have a larger mean than the advanced class.) The observed value of the test statistic is:

$$z = 0.67/0.427 = 1.57.$$

The P-value is 0.0582.

Chapter 8: Two "good" screening tests for a rare condition

The first screening test is described below.

	B	B ^c	Total
A	24,750	250	25,000
A ^c	2,499,750	247,475,250	249,975,000
Total	2,524,500	247,475,500	250,000,000

$$P(A) = 25,000/250,000,000 = 0.0001$$

$$P(B^c|A) = 250/25,000 = 0.01$$

$$P(B|A^c) = 2,499,750/249,975,000 = 0.01.$$

2,500,000 mistakes, of which 250 are false negatives

$$P(A^c|B) = 2,499,750/2,524,500 = 0.990$$

The second screening test is described below.

	B	B ^c	Total
A	24,999	1	25,000
A ^c	9,999	249,965,001	249,975,000
Total	34,998	249,965,002	250,000,000

$$P(A) = 25,000/250,000,000 = 0.0001$$

$$P(B^c|A) = 1/25,000 = 0.00004$$

$$P(B|A^c) = 9,999/249,975,000 = 0.00004.$$

10,000 mistakes, of which 1 is a false negatives

$$P(A^c|B) = 9,999/34,998 = 0.286.$$