CHAPTER NINE: HYPOTHESIS TESTING

Review May 11 ❧ Test May 18

Research requires an understanding of underlying mathematical distributions as well as of the research methods concepts discussed in chapter 7. Spreadsheets allow large amounts of data to be used and analyzed easily, allowing for richer and more applicable results.

9-A Normal Distributions — Wednesday • 4/26

normal distribution • standardized score (z score)

❶ Convert a raw score x to a standardized score z.

❷ Find the area under the normal curve between 0 and z.

❸ Calculate normal probabilities or percentages from raw scores.

9-B The Central Limit Theorem — Friday • 4/28

law of large numbers • sampling distribution • standard error • central limit theorem

❶ Use the law of large numbers to determine whether a given sample statistic is more likely to occur in a small sample or a large sample.

❷ Calculate normal probabilities or percentages for samples.

9-C *P* Values — Tuesday • 5/2

p value

❶ Calculate and interpret a p value for a simple event, and relate it to a null hypothesis.

❷ Use a p value to make a hypothesis or statistical conclusion if appropriate.

❸ Use a p value to make a statistical conclusion about a test.

❹ Do a z test for a within-participants design.

9-D Types of Statistical Tests — Friday • 5/5

❶ Select an appropriate statistical test for a research hypothesis.

❷ Use a calculator to do a statistical test.

9-E Spreadsheet Data Analysis — Tuesday • 5/9

index

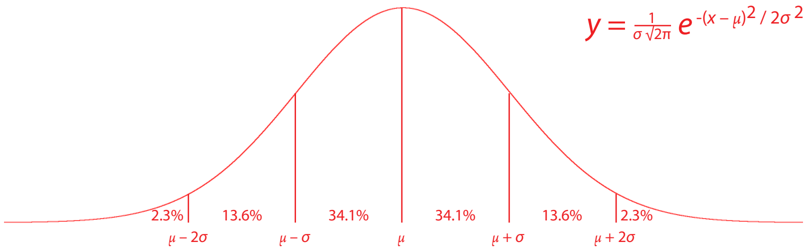
❶ Calculate values of an index as an operational definition of a conceptual variable.

❷ Use a spreadsheet to do a t test of two means or of a mean difference.

❸ Use a spreadsheet to do an r test of a correlation.

9-A Normal Distributions

The NORMAL Curve at right shows the NORMAL Probability Distribution. As shown by the curve, most data in a normal distribution are close to the mean, and very few are more than two standard deviations away from the mean. The curve is called normal because real-world data normally are distributed in this manner.



The percentages show the area under the curve in a given region, which is the same as the probability of a data value falling within that region.

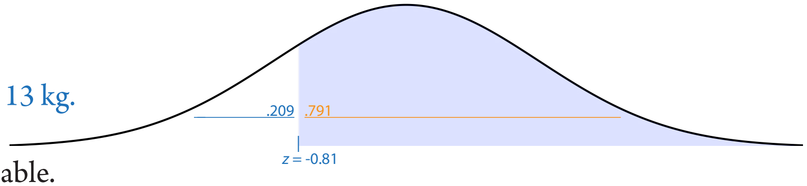
A STANDARDIZED Score, represented by z, is the number of standard deviations a raw score x is above the mean : $z = \frac{x - \mu}{\sigma}$. A Standard Normal Curve is a normal curve using standardized units, that is, z scores.

❶ Convert a raw score x to a standardized score z.

1. Subtract the mean from the raw score.
2. Divide the difference by the standard deviation.

❶ Calculate the z score a 25-year-old man weighing 75 kg, given $\mu = 77$ kg and $\sigma = 13$ kg.

$z = \frac{75-77}{13} \approx -0.15$



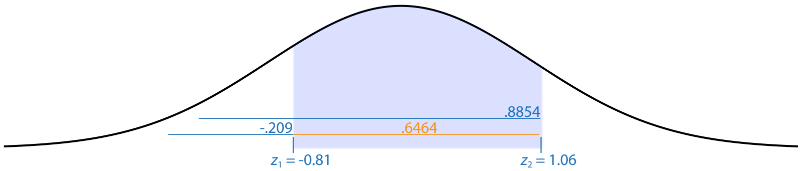
The area under a standard normal curve below a given z value can be looked up in a z table.

❷ Find the area under the normal curve between two z scores.

1 Use a z table to find the area under the normal curve below the first z score.
2. Use a z table to find the area under the normal curve below the second z score.
3. Subtract the smaller area from the larger area.

❷ Find the area under the normal cuve between -0.81 and 1.06.

1. $P(z < -0.81) = .209$
2. $P(z < 1.06) = .8554$
3. $P(-0.81 < z < 1.06) = .8554 - .2090 = .6464$



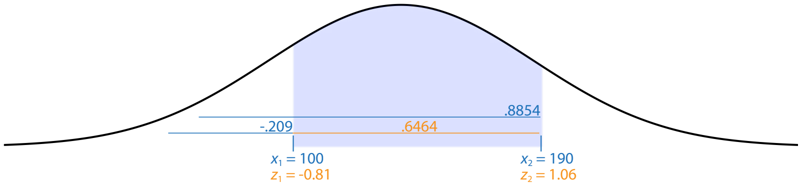
The z table only uses z scores, but raw scores (that is, scores using the original units) can be converted to z scores using the z formula above.

❸ Calculate normal probabilities or frequencies from raw scores.

1. Convert the raw scores to z scores (see ❶).
2. Use the z scores to calculate the probability (see ❷).

❸ On a test with normally distributed scores with $\mu = 139$ and $\sigma = 48$, what percent of scores are between 100 and 190?

1. $z_1 = \frac{100-139}{48} \approx -0.81$ $z_2 = \frac{190-139}{48} \approx 1.06$
2. $P(100 < x < 190) = P(-0.81 < z < 1.06) \approx .8554 - .2090 \approx 64.6\%$



9-B The Central Limit Theorem

The LAW OF LARGE NUMBERS states that the larger a sample is, the more accurately statistics calculated from the sample tend to represent the actual population parameters. As a result, larger samples are more likely to have results that turn out about as expected.

❶ Use the law of large numbers to determine whether a statistic is more likely to fall within a given range for a small sample or for a large sample.

1. Identify whether or not the range includes the population parameter.
2. If so, the larger sample is more likely to include the stated statistic. Otherwise, the smaller sample is more likely.

❶ Is Trump more likely to have a favorable approval rating in a random sample of 20 California voters or in a random sample of 200 California voters?

1. Trump's favorability rating is low in California. It is clearly not in the range $p > 50\%$.
2. He is not likely to have a favorable rating in either sample, but it could happen by coincidence in either sample. Coincidences are more likely in smaller samples, which in this case is the sample of 20 California voters.

Statisticians take advantage of the law of large numbers with SAMPLING Distributions, which are distributions of an entire statistic. In particular, distributions of sample means are frequently used.

On a normal curve, the probability of a single data value being at least one standard deviation above the mean is relatively low at 16%, but, by the law of large numbers, the probability of two data values averaging at least one standard deviation above the mean is even lower at 8%. This can be explained by the fact that both values would have to be extreme for their average to be extreme. In other words, extreme averages are less common than extreme individual values. The greater the number of values being averaged, the less likely an extreme value is. Therefore, the standard deviation of the sample means decreases as the sample size being used gets larger.

The standard deviation of sample means is labeled $\sigma_{\bar{x}}$ (or $s_{\bar{x}}$) and is called the STANDARD ERROR of the Mean. The CENTRAL LIMIT Theorem states that the larger the sample size n is, and the more normal the x distribution is, the closer to normal the $\bar{x}$ distribution will be, with a standard error of $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$, especially if the distribution is somewhat symmetrical and $n \geq 30$. Therefore, for samples, the z formula becomes $z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$.

The central limit theorem is important in two respects. It quantifies the effect of the law of large numbers by providing a formula for the standard deviation of sample means. It also allows use of the normal curve even for distributions that are not clearly normal, since sample averages from the distribution do tend to be normal even if the individual scores are not.

❷ Calculate normal probabilities or percentages for samples.

1. Follow the steps in 9-A ❸, but divide the standard deviation by the square root of the sample size.

❷ On a test with normally distributed scores with $\mu = 140$ and $\sigma = 55$, what percent of random 4-score averages are between 195 and 230?

1. $z_1 = \frac{195 - 140}{55 / \sqrt{4}} \approx 2.00$ $\qquad z_2 = \frac{230 - 140}{55 / \sqrt{4}} \approx 3.27$
2. $P(195 < \bar{x} < 230) = P(2.00 < z < 3.27) = .4995 - .9772 = 2.23\%$

9-C *P* Values

The *P* VALUE of a predicted sample result is the probability of reaching the same result in another random sample if the null hypothesis is actually true. *P* is a conditional probability: It is the probability of a predicted event occuring, given the null hypothesis is true, which is not the same as the probability of the null hypothesis being true, given the predicted occuring. As an analogy, the probability of a random criminal being an adult is very different from the probability of a random adult being a criminal.

❶ Calculate and interpret a *p* value for a simple event, and relate it to a null hypothesis.

1. Calculate p = the probability of the reaching the same result again in another random sample of the same size, given there is no reason for this other than coincidence.
2. The p value is the probability of reaching these results, given the null hypothesis is true.
3. The p value is not the probability that the null hypothesis is true. This value is unknown, but is lower for lower p values.

❶ Matt predicts that four coins will all land on tails, and they do.

1. $p = (½)^4 \approx 6\%$
2. If Matt's prediction was correct coincidentally, there is a 6% chance it will turn out that way on the next attempt as well.
3. The probability that this was not a coincidence is not 6%, but rather is an unknown value.

❷ Use a *p* value to make a hypothesis or statistical conclusion if appropriate.

1. If p is not very low, such a result could easily be a coincidence, so do not draw any conclusions.
2. If p is very low, consider reasons other than coincidence that could explain the result:
   a) If there are no reasonable explanations, the p value is meaningless and no conclusions should be drawn.
   b) If there is reasonable explanation that was not considered until after the results were known, the p value itself is meaningless and no conclusions should be drawn, but a hypothesis can be formed and tested with a new sample. It would not be logical to use the original data to make a prediction about themselves.
   c) If there is a reasonable explanation that was hypothesized beforehand, conclude that this hypothesis is statistically valid.

❷ The average semester grade in Statistics last spring was $\bar{x}_1 = 87.8\%$ with $s_1 = 7.96\%$ for the 31 girls and $\bar{x}_2 = 78.0\%$ with $s_2 = 8.19\%$ for the 38 boys. The p value for these statistics is $p = .000002$.

2. This is an extremely small value.

a) If there is no reason, other than this result, to expect girls to continue outscoring boys, then there is no significance to this result. It is likely coincidental.

b) If a reasonable explanation can be developed after the fact, such as teenage girls being more organized and studious than teenage boys, then this hypothesis can be tested on a new sample, such as this spring's Statistics grades. No conclusion can be made from the original data, however.

c) If a reasonable explanation, such as the one above or past observations, had been given beforehand as a basis for hypothesizing this result, then it is appropriate to conclude that girls do in fact outscore boys in this context. It is possible, however, that this conclusion is incorrect; the probability of this is not .000002.

Formally, the standard in scientific hypothesis testing is to define “very low” for a p value as anything below 5%, and the conclusion from such a low p value is to reject H_0. Therefore, if the p value of a predicted event is below 5%, there is sufficient evidence to claim that the reason the results were different from H_0 was not coincidence but rather because H_0 is actually false. It is important to keep in mind that any such claim may be a type I error, and, conversely, that failing to make such a claim may be a type II error.

❸ Use a p value to make a statistical conclusion about a test.

1. If $p < .05$, claim that H_0 is false, and specify whether you are claiming that the true population parameter is higher or lower than stated by H_0. However, be aware that you may be making a type I error.
2. If $p > .05$, do not make a claim about H_0. However, be aware that you may be making a type II error.
3. Be sure never to claim that H_0 is true. You either have sufficient evidence to claim that it is false, or you do not.
4. Be sure never to claim anything other than what was predicted.

❸ Calder hypothesizes that there are more Democrats than Republicans in Scotts Valley. In his random sample, there are 18 Democrats and 10 Republicans. He calculates $p = .065$.

2. Although Calder’s results turned out as predicted—there were more Democrats than Republicans in his sample (almost double, in fact)—he does not have sufficient evidence to conclude that there are more Democrats than Republicans in the overall Scotts Valley population. He may be making a type II error.

In a within-participants design, participants' data values are not actually used; instead, each participant's difference is used. For example, the value for a participant scoring 20 with eyes open and 12 with eyes closed would be 8, the same as a for a participant scoring 58 with eyes open and 50 with eyes closed. H_0 generally involves there being no difference between conditions, that is, the mean difference is $\mu = 0$.

In order to do a z test, the population standard deviation σ must be known. If it is not known, a similar distribution called Student's t distribution can be used (see 9-D), but for now we will use the normal distribution and use the sample standard deviation s to estimate the population standard deviation σ.

❹ Do a z test for a within-participants design.

1. State and justify the prediction.
2. For each participant, find the difference between the two trials. Do each subtraction in the same direction, such as version A minus version B, even if this results in negatives for some differences.
3. Calculate the sample mean and sample standard deviation of the differences. This can be done by hand (see 7-B ❺), on the calculator (see Statistics & Research Methods 3-B ❸), or in a spreadsheet (see 8-A).
4. Calculate z (see 9-B), using the sample mean for $\bar{x}$, 0 for μ, s as an estimate for σ, and the number of participants for n.
5. Use a normal table to calculate p (see 9-B ❷).
6. State the statistical conclusion in a sentence that makes the context clear (see ❷), followed by the calculated statistic z and the p value if H_0 is rejected or "$p > .05$" if H_0 is not rejected.

❹ Mitchell is testing the idea of interference limiting memory. He gives participants a minute to memorize a list of 20 words before giving them a recall test on these words. He then repeats this with a new list of 20 words.

1. He predicts that people will do worse the second time because the words on the first list are interfering with memory of the second list.

Participant #:	1	2	3	4	5	6	7	8
# correct in round one	12	14	13	18	11	19	16	15
# correct in round two	10	11	13	12	13	15	12	14
2. Difference:	2	3	0	6	-2	4	4	1

3. $\bar{x} = 2.25$, $s = 2.55$

4. $z = \frac{2.25 - 0}{2.55 / \sqrt{8}} = 2.50$

5. $p = P(\bar{x} > 2.25) = P(z > 2.50) = 1 - .994 = .006$

6. $p < .05$, so reject H_0: People have a harder time remembering a list of words if they have memorized a different list just beforehand, $z = 2.50$, $p = .006$.

9-D Types of Statistical Tests

Not all research hypotheses can be tested with a z test of a single mean. Below are some common types of statistical tests.

Type of test	Statistic	What is being tested	Example
single mean	z or t	Is a population mean different from a specified value?	Are cross country runners' resting heart rates lower than 60 beats per minute on average?
single mean difference	z or t	Is one population mean different from another in a within-participants design?	Are cross country runners' resting heart rates lower at the end of the season than at the beginning?
two means	z or t	Is one population mean different from another in a between-participants design?	Are cross country runners' resting heart rates lower than soccer players' resting heart rates?
ANOVA	F	Are multiple population means not all the same?	Does resting heart rate vary between cross country runners, soccer players, and basketball players?
single proportion	z	Is a population proportion different from a specified value?	Are more than 25% of cross country runners freshmen?
two proportions	z	Is one population proportion different from another?	Is the proportion of cross country runners who are freshmen higher than the proportion of soccer players who are freshmen?
goodness of fit	χ^2	Is a population distribution different from a specified distribution?	Is the distribution of cross country runners by grade level not uniform (that is, not 25% in each grade)?
single variance	χ^2	Is a population standard deviation different from a specified value?	Is the standard deviation of boys' times at CCS finals greater than 90 seconds?
two variances	F	Is one population standard deviation different from another?	Is the standard deviation of times at CCS finals different in the boys race than in the girls race?
correlation	r	Is one numerical variable correlated with another?	Do racers tend to run faster the more miles they ran in the summer?
independence	χ^2	Is one categorical variable correlated with another?	Does favorite subject vary by sport?

❶ Select an appropriate statistical test for a research hypothesis.

1. Use the chart above to see what type of test matches the research hypothesis, or use the flowchart in Statistics & Research Methods 9-G. Some complicated research hypotheses do not fit into the above categories.

❶ Do high school students spend more time on their phones than college students do?

There are two variables: school and phone time.

School is not continuous because it only has two possible values in this study, high school and college.

Average (mean) phone time is being compared.

It is a between-participants design, as each participant is in high school or college but not both.

There is no way to know the actual population standard deviations.

t test of two means

Most of the tests listed above can be done on a graphing calculator.

❷ Use a calculator to do a statistical test.

1. Identify the type of test (see ❶).
2. View the specified section in the Statistics notes for step-by-step directions, or search for directions from other online sources.

❷ Does number of absences help predict semester grade in math? The current grade and number of spring semester absences for January through April of eight random Statistics students are shown below.

# of absences	4	3	8	5	2	2	7	3	8	2
current grade	94	90	72	76	89	83	82	82	95	74

1. This is an *r* test of a correlation, because one numerical variable (number of absences) is hypothesized to correlate with another numerical variable (semester grade in math).
2. Calculator directions are shown in Statistics & Research Methods 9-A ❸.

   The test is left-tailed, because we are predicting that a higher number of absences predicts a lower grade.

   $p = .45$, so do not reject H_0. We do not have sufficient evidence to conclude that the more absences a student has, the lower their Statistics grade is likely to be, $r = -.05$, $p > .05$.

9-E Spreadsheet Data Analysis

Spreadsheets make it easy to analyze large amounts of data.

An INDEX is a single variable compiled from multiple variables and is used as an operational definition for a conceptual variable. For example, a heath index might incorporate blood pressure, dietary habits, and exercise habits. Each variable is mathematically factored in. The variables can all be weighted equally, or can have different weightings based on how important they are to the conceptual variable being indexed. They do not have to be weighted proportionally; for example, hours of exercise per week may have a log function or a root function applied, making the difference between no exercise and one hour per week much more significant than the difference between 20 hours and 21 hours.

❶ Calculate values of an index as an operational definition of a conceptual variable.

1. Choose measurable factors that would appropriately represent the conceptual variable.
2. Decide if any of the factors are more important than others.
3. Make a column for each factor and a column for the index variable.
4. Below each factor enter a weighting (optional).
5. In the index variable column, make a formula that takes each factor into account. If you did step 4, reference these cells in the formula (using $ to lock their location in row 2) rather than typing in constant weightings.
6. Copy the formula down the column so that it will automatically be calculated for each new entry.

❶ Use an index to rate applicants for a leadership scholarship.

1. application essay (score of 0 to 10)
   letter of recommendation (score of 0 to 10)
   currently in student government (yes or no)
2. The application essay will be weighted most heavily.

	A	B	C	D	E
1	Applicant #	Essay Score	Letter of Rec Score	Student Government	Calculated Rating
2	Weighting:	50%	30%	20%	
3	1	9	6	no	=(B$2*B3+C$2*C3)*(1+if(D3=”yes”,1+D$2,1))
4	2	7	7	yes	=(B$2*B4+C$2*C4)*(1+if(D4=”yes”,1+D$2,1))
5	3	8	10	yes	=(B$2*B5+C$2*C5)*(1+if(D5=”yes”,1+D$2,1))

T tests can be done in a spreadsheet. The procedure shown below is based on statistical concepts that are beyond the scope of this course.

❷ Use a spreadsheet to do a *t* test of two means or of a mean difference.

1. Enter the data from one sample into a column (`range1`) and the data from the other sample in another column (`range2`). For a test of a mean difference, make sure each participant's two data values are in the same row.
2. Calculate the mean of each sample: $\bar{x}_1$ `=AVERAGE(range1)`, $\bar{x}_2$ `=AVERAGE(range2)`
3. Calculate the standard deviation for each sample: s_1 `=STDEV(range1)`, s_2 `=STDEV(range2)`
4. Count the sample sizes: n_1 `=COUNT(range1)`, n_2 `=COUNT(range2)`
5. Calculate the degrees of freedom: df_1 `=n1-1`; if between participants, then also df_2 `=n2-1`
6. Calculate the *p* value: *p* `=TTEST(range1, range2, tails, type)`, where `tails` is the number of tails (`1` or `2`) and `type` is `1` for within-participants (mean difference) or `2` for between-participants (two means).
7. Calculate the *t* score: *t* `=-t.inv(p,df)` for one-tailed, or `=tinv(p,df)` for two tailed
8. Use the *p* value to make a statistical conclusion about the test (see 9-C ❷).
9. State the conclusion, followed by $t(df)$, *p*.

❷ From a sample of 11 students in the same math class, 5 are randomly assigned to review class notes for at least 10 minutes each night after class throughout one chapter. The scores for that chapter's test are compared.

	A	B	C	D
1	Condition:	Study	Control	
2		88	92	
3		84	71	
4		94	90	
5		96	76	
6		91	82	
7			81	
8	Mean:	=AVERAGE(B2:B7)	=AVERAGE(C2:C7)	
9	Standard Deviation:	=STDEV(B2:B7)	=STDEV(C2:C7)	
10	Sample Size:	=COUNT(B2:B7)	=COUNT(C2:C7)	
11	Degrees of Freedom:	=B10-1	=C10-1	
12	*p*:			=TTEST(B2:B7,C2:C7,1,2)
13	*t*:			=-T.INV(D12,B11+C11)

Reviewing class notes for 10 minutes each night after class improves test scores, $t(9) = 2.10$, $p = .033$.

One of the most commonly used statistical tests is an *r* test of a correlation. Sheets does not have a function for this, but *r* can be converted to *t* to achieve the same result.

❸ Use a spreadsheet to do an *r* test of a correlation.

1. Enter the data for one variable into a column (`range1`) and the data for the other variable into another column (`range2`) so that each participant's two data values are in the same row.
2. Count the sample size: *n* `=COUNT(range1)`
3. Calculate the degrees of freedom: *df* `=n-2`
4. Calculate the correlation coefficient, which indicates how strongly, on a scale of 0 to 1, the two variables are correlated, and in which direction, positive or negative: *r* `=CORREL(range1,range2)`
5. Calculate the *p* value: *p* `=TDIST(SQRT((df*r^2)/(1-r^2)),df,tails)`
6. Use the *p* value to make a statistical conclusion about the test (see 9-C ❷).
7. State the conclusion, followed by *r*, *p*.

❸ The literacy rates (x) and life expectancies (y) are shown below for nine random nations.

	A	B	C	D
1	Country	Literacy Rate (%)	Life Expectancy (years)	
2	Argentina	97	75	
3	Brazil	89	72	
4	China	92	73	
5	Colombia	90	73	
6	Ecuador	91	75	
7	Egypt	72	73	
8	India	74	64	
9	Iran	85	72	
10	Pakistan	55	65	
11	Sample Size:			`=COUNT(B2:B10)`
12	Degrees of Freedom:			`=D11-2`
13	*r*:			`=CORREL(B2:B10,C2:C10)`
14	*p*:			`=TDIST(sqrt((D12*D13^2)/(1-D13^2)),D12,1)`

Life expectancy is positively correlated with literacy, that is, the higher a nation's literacy rate is, the longer its citizens are expected to live, $r = .79$, $p = .006$.