



1 — Estimation

1.1 Confidence Intervals

Definition 1.1 — Margin of error. The margin of error of a distribution is the amount of error we predict when estimating the population parameters from sample statistics. The margin of error is computed as:

$$Z^* \cdot \frac{\sigma}{\sqrt{n}}$$

Where Z^* is the critical z-score for the level of confidence.

Definition 1.2 — Confidence level. The confidence level of an estimate is the percent of all possible sample means that fall within a margin of error of our estimate. That is to say that we are some % sure the the true population parameter falls within a specific range

Definition 1.3 — Confidence Interval. A confidence interval is a range of values in which we suspect the population parameter lies between. To compute the confidence interval we use the formula:

$$\bar{x} \pm Z^* \cdot \frac{\sigma}{\sqrt{n}}$$

This gives us an upper and lower bound that capture our population mean.

1.1.1 Critical Values

The critical z-score is used to define a critical region for our confidence interval. Observations beyond this critical region are considered observations so extreme that they were very unlikely to have just happened by chance.

1.2 Practice Problems

Problem 1.1 Find a confidence interval for the distribution of pizza delivery times.

Company A
20.4
24.2
15.4
21.4
20.2
18.5
21.5

Table 1.1: Pizza Companies Delivery Times

What is a Hypothesis test?

Error Types

Practice Problems



2 — Hypothesis testing

2.1 What is a Hypothesis test?

A hypothesis test is used to test a claim that someone has about how an observation may be different from the known population parameter.

Definition 2.1 — Alpha level (α). The alpha level (α) of a hypothesis test helps us determine the critical region of a distribution.

Definition 2.2 — Null Hypothesis. The null hypothesis is always an equality. It is a the claim we are trying to provide evidence against. We commonly write the null hypothesis as one of the following:

$$H_0 : \mu_0 = \mu$$

$$H_0 : \mu_0 \geq \mu$$

$$H_0 : \mu_0 \leq \mu$$

Definition 2.3 — Alternative Hypothesis. The Alternative hypothesis is result we are checking against the claim. This is always some kind of inequality. We commonly write the alternative hypothesis as one of the following:

$$H_a : \mu_a \neq \mu$$

$$H_a : \mu_a > \mu$$

$$H_a : \mu_a < \mu$$

■ **Example 2.1** A towns census from 2001 reported that the average age of people living there was 32.3 years with a standard deviation of 2.1 years. The town takes a sample of 25 people and finds there average age to be 38.4 years. Test the claim that the average age of people in the town has increased. (Use an α level of 0.05)

First lets define our hypotheses:

$$H_0 : \mu_0 = 32.3\text{years}$$

$$H_a : \mu_0 > 32.3\text{years}$$

Now lets identify the important information:

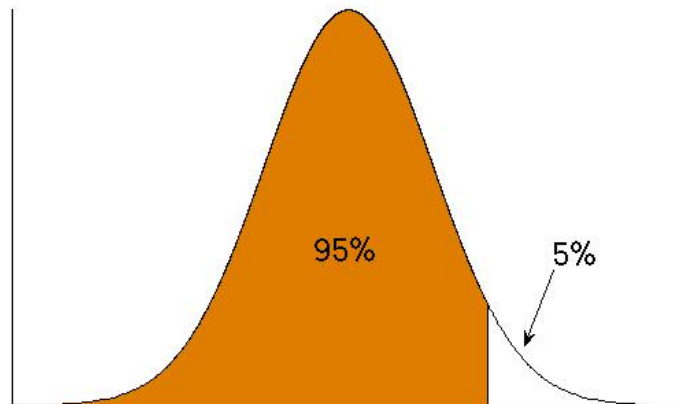
$$\bar{x} = 38.4$$

$$\sigma = 2.1$$

$$n = 25$$

$$SE = \frac{2.1}{\sqrt{25}} = 0.42$$

The last piece of important info we need is our critical value: Finding Z-critical value:



So we look up as close as we can to 95%

So that gives us a Z-crit of 1.64

Once we have all our important information we can now find our test statistic:

$$z\text{-score} = \frac{38.4 - 32.3}{0.42} = 14.5238$$

Since our z-score is much bigger than our z-crit we reject the claim (reject the null) that the average age of people living there was 32.3 years. ■

2.1.1 Error Types

Definition 2.4 — Type I Error. A Type I Error is when you reject the null when the null hypothesis is actually true. The probability of committing a Type I error is α

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857

Definition 2.5 — Type II Error. A Type II Error is when you fail to reject the null when it is actually false. The probability of committing a Type II error is β

2.2 Practice Problems

Problem 2.1 An insurance company is reviewing its current policy rates. When originally setting the rates they believed that the average claim amount was \$1,800. They are concerned that the true mean is actually higher than this, because they could potentially lose a lot of money. They randomly select 40 claims, and calculate a sample mean of \$1,950. Assuming that the standard deviation of claims is \$500, and set $\alpha = 0.05$, test to see if the insurance company should be concerned.

Problem 2.2 Explain a type I and type II error in context of the problem. Which is worse?



3 — t-Tests

3.1 t-distribution

The t-Test is best to use when we do not know the population standard deviation. Instead we use the sample standard deviation.

Definition 3.1 — t-stat. The t-Test statistic can be computed very similarly to the z-stat, to compute the t-stat we compute:

$$t = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

We also have to compute the degrees of freedom (df) for the sample: $df = n - 1$

Like the Z-stat we can use a table to get the proportion below or between a specific value. T-tests are also great for testing two sample means (i.e. paired t-tests), we modify the formula to become:

$$\frac{(x_2 - x_1) - (\mu_2 - \mu_1)}{\frac{\sqrt{(s_1^2 + s_2^2)}}{n}}$$

■ **Example 3.1** ■

3.1.1 Cohen's d

Definition 3.2 — Cohen's d. Cohen's d measures the effect size of the strength of a phenomenon. Cohen's d gives us the distance between means in standardized units. Cohen's d is computed by:

$$d = \frac{\bar{x}_1 - \bar{x}_2}{s}$$

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

3.2 Practice Problem

Problem 3.1 Pizza company A wants to know if they deliver Pizza faster than Company B. The following table outlines there delivery times:

Company A	Company B
20.4	20.2
24.2	16.9
15.4	18.5
21.4	17.3
20.2	20.5
18.5	
21.5	

Table 3.1: Pizza Companies Delivery Times

Problem 3.2 Use Cohen's d to measure the effect size between the two times.



4 — t-Tests continued

4.1 Standard Error

Definition 4.1 — Standard Error. The Standard error is the standard deviation of the sample means over all possible samples (of a given size) drawn from the population. It can be computed by:

$$SE = \frac{\sigma}{\sqrt{n}}$$

The standard error for two samples can be computed with:

$$SE = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

Definition 4.2 — Pooled Variance. Pooled variance is a method for estimating variance given several different samples taken in different circumstances where the mean may vary between samples but the true variance is assumed to remain the same. The pooled variance is computed by using:

$$S_p^2 = \frac{SS_1 + SS_2}{df_1 + df_2}$$

R We can use pooled variance to compute standard error that is:

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



5 — One-way ANOVA

5.1 Anova Testing

The grand mean of several data sets is simply the sum of all the data divided by the number of data points. The grand mean is commonly given the symbol \bar{x}_G

Definition 5.1 — Between-Group Variability. Describes the distance between the sample means of several data sets and can be computed as the Sum of Squares Between divided by the degrees of freedom between:

$$SS_{between} = n \sum (\bar{x}_k - \bar{x}_G)^2$$

$$df_{between} = k - 1$$

where k is the number samples

Definition 5.2 — Within-Group Variability. Describes the variability of each individual sample and can be computed as the Sum of Squares within divided by the degrees of freedom within:

$$SS_{within} = \sum (x_i - \bar{x}_k)^2$$

$$df_{within} = N - k$$

The hypotheses for a typical anova test are:

$$H_0 : \mu_0 = \mu_1 = \dots = \mu_k$$

H_a : any of these means differs

5.1.1 F-Ratio

The F-ratio can be found by taking the between-group variability and dividing by the within-group variability. The F-ratio is used in the same way as the t-stat, or z-stat.

5.2 Practice Problem

Problem 5.1 Neuroscience researchers examined the impact of environment on rat development. Rats were randomly assigned to be raised in one of the four following test conditions: Impoverished (wire mesh cage - housed alone), standard (cage with other rats), enriched (cage with other rats and toys), super enriched (cage with rats and toys changes on a periodic basis). After two months, the rats were tested on a variety of learning measures (including the number of trials to learn a maze to a three perfect trial criteria), and several neurological measure (overall cortical weight, degree of dendritic branching, etc.). The data for the maze task is below. Compute the appropriate test for the data provided below. What would be the null hypothesis in this

Impoverished	Standard
Enriched	Super Enriched
22	17
12	8
19	21
14	7
15	15
11	10
24	12
9	9
18	19
15	12

Table 5.1: Scores

study

Problem 5.2 What is your F-critical?

Problem 5.3 What is your F-stat?

Problem 5.4 Are there any significant differences between the four testing conditions?



6 — ANOVA continued

6.1 Means

Definition 6.1 — Group Means. The group means are the individual mean for each group in an Anova test.

Definition 6.2 — Mean Squares. The $MS_{between}$ and MS_{within} are computed as:

$$MS_{between} = \frac{SS_{between}}{df_{between}}$$

$$MS_{within} = \frac{SS_{within}}{df_{within}}$$

6.2 Tukey's HSD

Definition 6.3 — Tukey's HSD. Tukey's HSD allows us to make pairwise comparisons to determine if a significant difference occurs between means. If Tukey's HSD is greater than the difference between sample means then we consider the samples significantly different. Keep in mind that the sample sizes must be equal. Tukey's HSD is computed as:

$$q^* \sqrt{\frac{MS_{within}}{n}}$$

We can also use Cohen's d for multiple comparisons on sample sets. Using Cohen's d we have to compute the value for every possible combination of samples.

Definition 6.4 — η^2 . η^2 (read eta squared) is the proportion of total variation that is due to between group differences.

$$\eta^2 = \frac{SS_{between}}{SS_{within} + SS_{between}} = \frac{SS_{between}}{SS_{total}}$$

R The value of η^2 is considered large if it is greater than 0.14

6.3 Practice Problems

Problem 6.1 Amy is trying to set-up a home business of selling fresh eggs. In order to increase her profits, she wants to only use the breed of hens that produce the most eggs. She decides to run an experiment testing four different breeds of hens, counting the number of eggs laid by each breed. She purchases 10 hens of each breed for her experiment. What is the studentized range statistic (q^*) for this experiment at an alpha level of 0.05?

Problem 6.2 Amy finds that the MS_{within} for the first batch of eggs laid by her hens to be 45.25. How far apart do the group means for the different breeds have to be to be considered significant?

Problem 6.3 Amy also finds that $SS_{within} = 1629.36$ and $SS_{between} = 254.64$. What proportion of the total variation in the number of eggs produced by each breed can be attributed to the different breeds? (Calculate eta-squared)

Problem 6.4 Using Tukey HSD, are the sample means significantly different?



7 — Correlation


7.1 Scatterplots

A scatterplot shows the relationship between two sets of data. Each pair of data points is represented as a single point on the plane. The more linear our set of points are the stronger the relationship between the two data sets is.

7.1.1 Relationships in Data

Definition 7.1 — Correlation coefficient (Pearson's r). The Correlation coefficient, commonly referred to as Pearson's r, describes the strength of the relationship between two data sets. The closer $|r|$ is to 1 the more linear (stronger) our relationship. The closer r is to zero the more scattered (weaker) our relationship. To compute Pearson's r you can use the formula:

$$r = \frac{\text{Covariance}(x,y)}{S_x \cdot S_y}$$

 On a Google Docs spreadsheet we can do

```
=Pearson(start cell for variable x : end cell for variable x,  
start cell for variable y : end cell for variable y)
```

Definition 7.2 — Coefficient of Determination(r^2). The coefficient of determination is the percentage of variation in the dependent variable (y) that can be explained by variation in the independent variable (x)

7.2 Practice Problems

Problem 7.1 A researcher wants to investigate the relationship between outside temperature and the number of reported acts of violence. For this investigation, what is the predictor (x) variable and what is the outcome (y) variable?

Problem 7.2 Given a correlation coefficient of -0.95 , what direction is the relationship and how do we know this? What is the strength of this relationship and how do we know this? In terms of strength and relationship, how does this correlation coefficient differ from one that is 0.95 ?

Problem 7.3 What does it mean if we have a coefficient of determination = 0.55 ?

Problem 7.4 If a researcher found that there was a strong positive correlation between outside temperature and the number of reported acts of violence, does this mean that an increase or decrease in temperature causes an increase or decrease in the number of reported acts of violence? Why or why not?



8 — Regression

8.1 Linear Regression

Definition 8.1 — Regression Equation. The linear regression equation $\hat{y} = ax + b$ describes the linear equation that represents the "line of best fit". This line attempts to pass through as many of the points as possible. a is the slope of our linear regression equation and represents the rate of change in y versus x . b represents the y -intercept.

R The regression equation may also be written as $\hat{y} = bx + a$

The line of best first helps describe the dataset. It can also be used to make approximate predictions of how the data will behave.

Corollary 8.1 We can find the linear regression equation with the two following pieces of information:

$$\text{slope} = r \frac{s_y}{s_x}$$

The regression equation passes through the point (\bar{x}, \bar{y})

■ **Example 8.1** ■

8.2 Practice Problems

Problem 8.1 Marcus wants to investigate the relationship between hours of computer usage per day and number of minutes of migraines endured per day. After collecting data, He finds a correlation coefficient of 0.86, with $s_y = 375.55$ and $s_x = 1814.72$. The mean hours of computer usage from his data set was calculated to be 4.5 hours and the average number of minutes of migraine was calculated to be 25 minutes. Find the regression line that best fits his data.

Problem 8.2 Using the line that you calculated above, given 2 hours of computer usage, how many minutes of migraine would Marcus predict to follow?

Problem 8.3 Marcus coincidentally has a point in his data set that he collected for exactly 2 hours of computer usage. Given that the residual between his observed value for 2 hours of computer usage and the expected value (as calculated in the previous question) equals 1.89, how many minutes of migraine did Marcus observe for that point in his data set?

Scales of measurement

Chi-Square GOF test

Chi-Square test of independence

Practice Problem



9 — Chi-Squared tests

9.1 Scales of measurement

Definition 9.1 — Ordinal Data. There is a clear order in the data set but the distance between data points is unimportant.

Definition 9.2 — Interval Data. Similar to an ordinal set of data in that there is a clear ranking, but each group is divided into equal intervals

Definition 9.3 — Ratio Data. Similar to interval data except there exists an absolute zero.

Definition 9.4 — Nominal Data. This is the same as qualitative data, where we differentiate between items or subjects based only on their names and/or categories and other qualitative classifications they belong to.

Type of Data	Example	Data
Ordinal	Ranks in a race	1st, 2nd, 3rd
Interval	Temperature in Celsius	$-10^\circ - 0^\circ, 1^\circ - 10^\circ, 11^\circ - 20^\circ$
Ratio	Percentage correct on test	0 – 10%, 11 – 20%, 21 – 30%
Nominal	Shirt Colors	Red, Blue, Yellow, White

Table 9.1: Examples of different scales of measurement

■ **Example 9.1** ■

9.2 Chi-Square GOF test

The Chi-Square GOF test allows us to see how well observed values match expected values for a certain variable. In particular we compare the frequencies of our data sets.

9.2.1 Chi-Square test of independence

This variation of the Chi-Square test is used to determine if 2 nominal variables are independent. In particular we use the marginal totals.

9.3 Practice Problem

Problem 9.1 A poker-dealing machine is supposed to deal cards at random, as if from an infinite deck. In a test, you counted 1600 cards, and observed the following: table[h]

Suit	Count
Spades	404
Hearts	420
Diamonds	400
Clubs	376

Card counts

Could it be that the suits are equally likely? Or are these discrepancies too much to be random?