Statistics Exam - DONE

Recitation 1

Exercise 1

Given the following table

Student	Gender	Grade	City	Faculty	Family Income
1	F	28	Milan	Economics	18,000
2	F	25	Florence	Statistics	53,000
3	M	23	Naples	Political Sciences	100,000
4	F	30	Rome	Economics	13,000
5	M	25	Rome	Statistics	89,000
6	M	28	Milan	Statistics	89,000

1. Which are the subjects?
2. Which are the variables?
3. Which type of variables are they?

1. Students.
   The subjects are the "keys" of the table. If i'm not retarded they should always be unique.
2. Gender, Grade, City, Faculty, Family Income.
   The variables are just the fields of the table.
3. Variable types:
   1. Gender: categorical (Binary, with two categories M, F)
   2. Grade: quantitative discrete
   3. City: categorical (Nominal, with categories Milan, Rome, Naples, Florence)
   4. Faculty: categorical (Nominal, with categories Economics, Statistics, Political Sciences)
   5. Family Income: quantitative continuous

Brief recap on variable types

Categorical (or Qualitative) variable

Each observation belongs to a category:

- Binary: There are only 2 categories
- Ordinal: The categories have a hierarchy or order.
- Nominal: No hierarchy is present.

Quantitative variable

Observations take numerical values that represent different magnitudes of the variable:

- Discrete: The possible values come from a specific set of numbers
- Continuous: The values are from an interval.

Exercise 2

The following table reports the number of sick leaves for a sample of employees

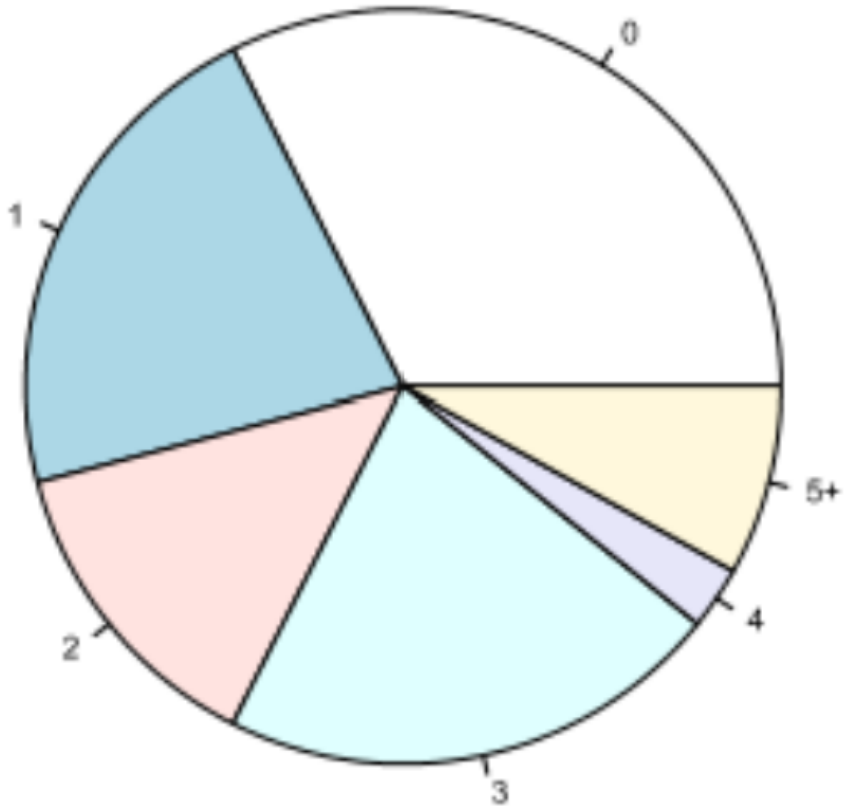
Value	Frequency
0	12
1	8
2	5
3	8
4	1
5+	3

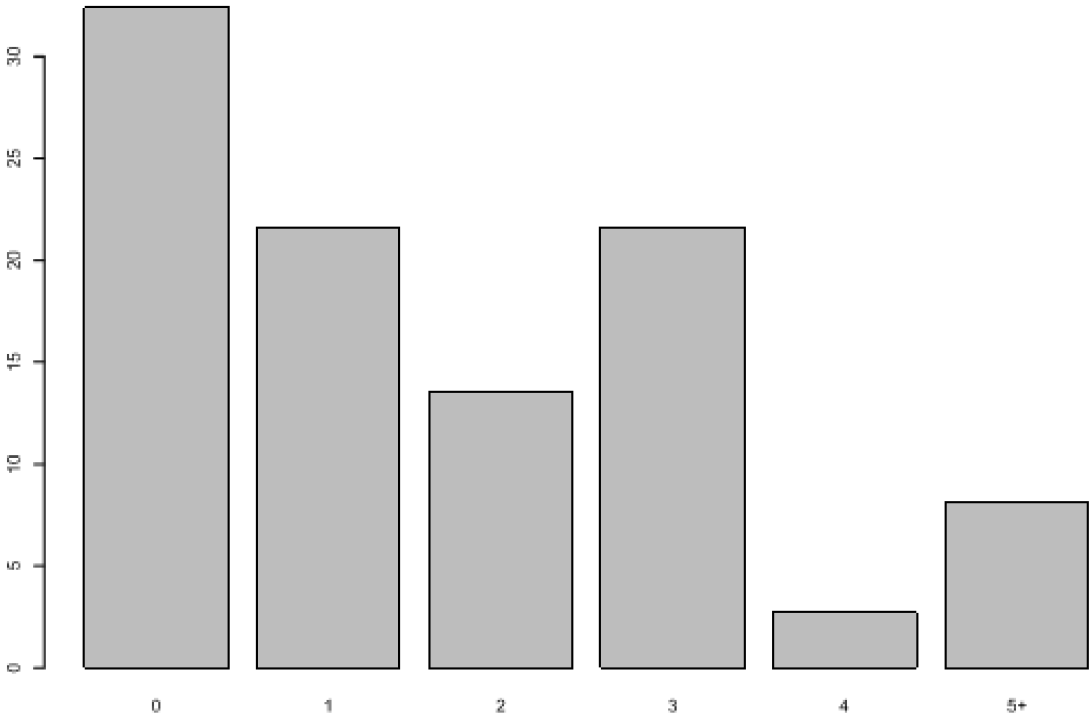
1. Add proportions and percentages to the above table
2. Draw a pie chart and a bar chart to represent these data

1. A proportion is just $\frac{Frequency}{Total\ Observations}$:

Value	Frequency	Proportions	Percentages
0	12	0.32	32
1	8	0.22	22
2	5	0.13	13
3	8	0.22	22
4	1	0.03	3
5	3	0.08	8
Total	37	1	100

2. Pretty simple innit:





Exercise 3

The following data represent the sizes of 20 families that live in a small town in Guatemala:

$$5, 7, 10, 8, 6, 6, 10, 7, 5, 10, 8, 11, 8, 6, 9, 5, 10, 7, 11, 5$$

1. Build the frequency table with counts, proportions and percentages

1. The subjects here is the number of people in the family, and the frequency is how many families have this number of people in it:

Value	Frequency	Proportions	Percentages
5	4	0.20	20
6	3	0.15	15
7	3	0.15	15
8	3	0.15	15
9	1	0.05	5
10	4	0.20	20
11	2	0.10	10
Total	20	1	100

Exercise 4

15 people attended the launch of a new book and the time required for reading the book (in months) for each of them is as follows

$$0.55, 0.61, 0.48, 0.64, 0.54, 0.43, 0.40, 0.42, 0.42, 0.38, 0.42, 0.39, 0.5, 0.53, 0.8$$

1. Build the frequency table with counts, proportions and percentages by considering the following classes:

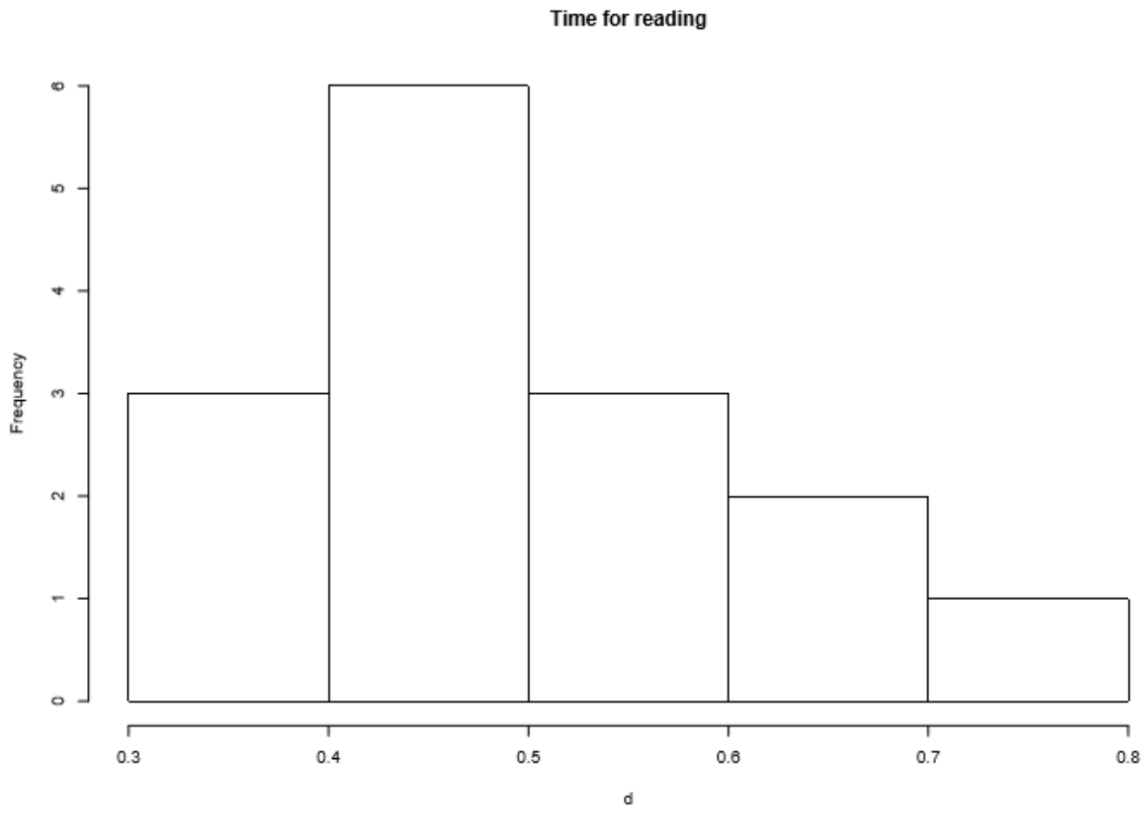
$$[0.30, 0.40), [0.40, 0.50), [0.50, 0.60), [0.60, 0.70), [0.70, 80]$$

2. Draw the histogram representing the distribution of the above variable
3. What can be said about the variable by looking at this graph?

1. This is getting boring:

Class	Frequency	Proportions	Percentages
$[0.3, 0.4)$	2	0.13	13
$[0.4, 0.5)$	6	0.40	40
$[0.5, 0.6)$	4	0.27	27
$[0.6, 0.7)$	2	0.13	13
$[0.7, 0.8]$	1	0.07	7
Total	15	1	100

2. An histogram is a chart used to display numerical data.:



Histogram vs Bar chart

The two main differences between a bar chart and a histogram are:

- The bar chart displays categorical discrete data
- The bars of the bar chart are not adjacent to each other, there is a slight padding.

3. The distribution is unimodal and skewed to the right.

Question

When we convert from continuous to intervals, what is the type of the new variable?

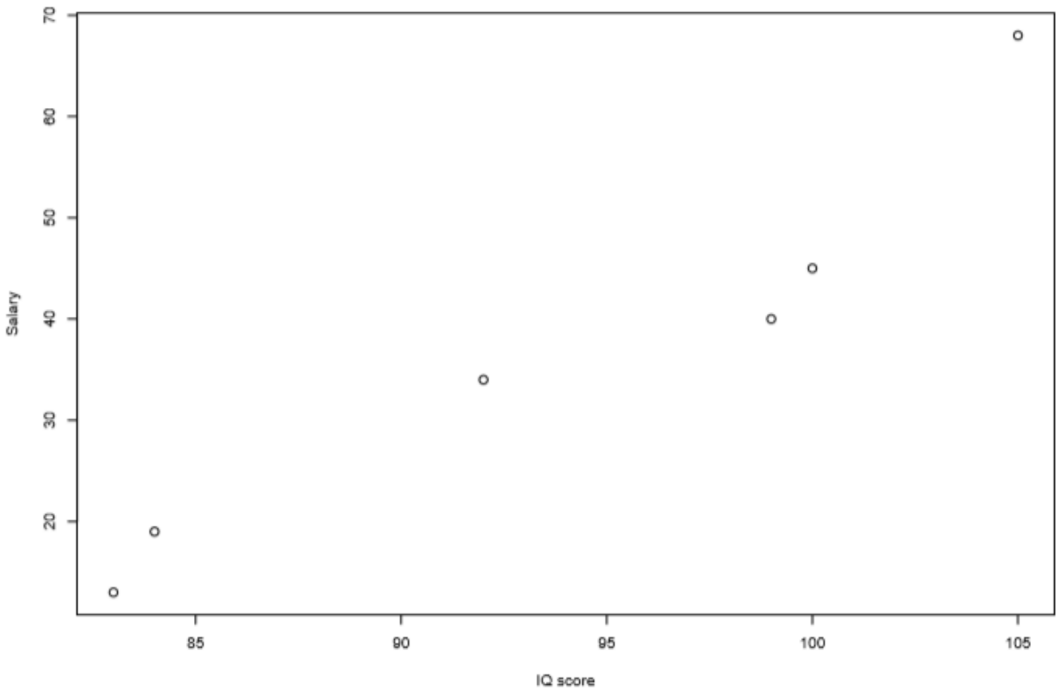
Exercise 5

The following table reports the IQ scores on an intelligence test and the annual salaries (in thousands of euro) of a sample of employees

Subject	IQ	Salary
1	105	68
2	84	19
3	83	13
4	92	34
5	99	40
6	100	45

1. Draw a scatterplot representing the Salary as a function of the IQ scores

1. Each row is a data point, we treat IQ as the X variable and Salary as the Y variable, as we want to look at the correlation between Salary and IQ.



Exercise 6

The number of rainy days that occurred each month are:

$$10, 14, 14, 8, 5, 6, 1, 3, 7, 10, 14, 9$$

1. Compute the mean, the median and the mode
2. Comparing the mean and the median, what can be said about the variable distribution?
3. If the number of rainy days doubles each month, what will be the mean? And the median? And the mode?

1. Mean, median and mode:
   1. Mean: $\frac{10+14+14+8+5+6+1+3+7+10+14+9}{12} = 8.41$
   2. Median:
      First we sort the data: 1, 3, 5, 6, 7, 8, 9, 10, 10, 14, 14, 14.
      The number of values is even, so we have 2 middle values, 8 and 9.
      We take the mean of these two numbers: $\frac{8+9}{2} = 8.5$.
   3. Mode: 14.
2. Median > Mean, therefore the distribution is slightly skewed to the left:

3. This question is so poorly written omfg, the new values are $1, 6, 10, 12, 14, 16, 18, 20, 20, 28, 28, 28$:
   1. Mean: $\frac{2*(20+28+28+16+10+12+1+6+14+20+28+18)}{12} = 16.8$
   2. Median: The new middle values are 16 and 9, so the new median is $\frac{16+18}{2} = 17$.
   3. Mode: 28.

Exercise 7

Consider a sample of 10 MBA graduates, whose first salary (in thousands of dollars per year) after graduation are as follows:

$$200, 70, 88, 130, 175, 89, 95, 120, 400, 55$$

1. Compute the mean and the median salary
2. Comparing the mean and the median, what can be said about the salary distribution?
3. What happens to the mean if the second year after graduation all salaries increase by 10 %?
4. What happens to the mean if all MBA graduates receive an extra amount of 10 thousand dollars?
5. What happens to the mean if all MBA graduates receive an extra amount of 10 thousand dollars in addition to an increase of 10%?
6. What happens to the mean and to the median if we accidentally report 4000 in place of 400?

1. Mean, and median:
   1. Mean: $\frac{200+70+88+130+175+89+95+120+400+55}{10} = 142.2$
   2. Median: $\frac{95+120}{2} = 107.5$
2. Skewed to the right, because Median < Mean.
3. New $\bar{x} = \frac{1.1*(200+70+88+130+175+89+95+120+400+55)}{10} = 1.1 * 142.2 = 156.4$
4. New $\bar{x} = \frac{(10*10)+(200+70+88+130+175+89+95+120+400+55)}{10} = 152.2$

5. New $\bar{x} = \frac{(10*10)+1.1*(200+70+88+130+175+89+95+120+400+55)}{10} = 166.4$
6. Change in mean and median:
    1. New $\bar{x} = \frac{200+70+88+130+175+89+95+120+4000+55}{10} = 502,2$
    2. Absolutely nothing lol. Because the order doesn't change and 400 wasn't in the middle anyways.

Recitation 2

Exercise 1

Consider a sample of 10 MBA graduates, whose first salary (in thousands of dollars per year) after graduation are as follows:

$$200, 70, 88, 130, 175, 89, 95, 120, 400, 55$$

1. Compute the sample standard deviation and the range
2. What happens to the sample standard deviation if the second year after graduation all salaries increase by 10 %?
3. What happens to the sample standard deviation if all MBA graduates receive an extra amount of 10 thousand dollars?
4. What happens to the sample standard deviation if all MBA graduates receive an extra amount of 10 thousand dollars in addition to an increase of 10%?

1. The sample standard deviation is computed as follows: $\sqrt{\frac{\sum(x_i-\bar{x})^2}{N-1}}$
2. We are gonna do this later.

Why N - 1?

This is a non-technical explanation

Because the sample standard deviation is computed as an approximation of the real standard deviation, from a sample of the population.

This means that the data point we get are more likely to be around the mean and less likely to be on the tails of the distribution.

So the sample standard deviation always underestimates the real value.

For this reason we decrease the denominator and overshoot the number.

Exercise 3

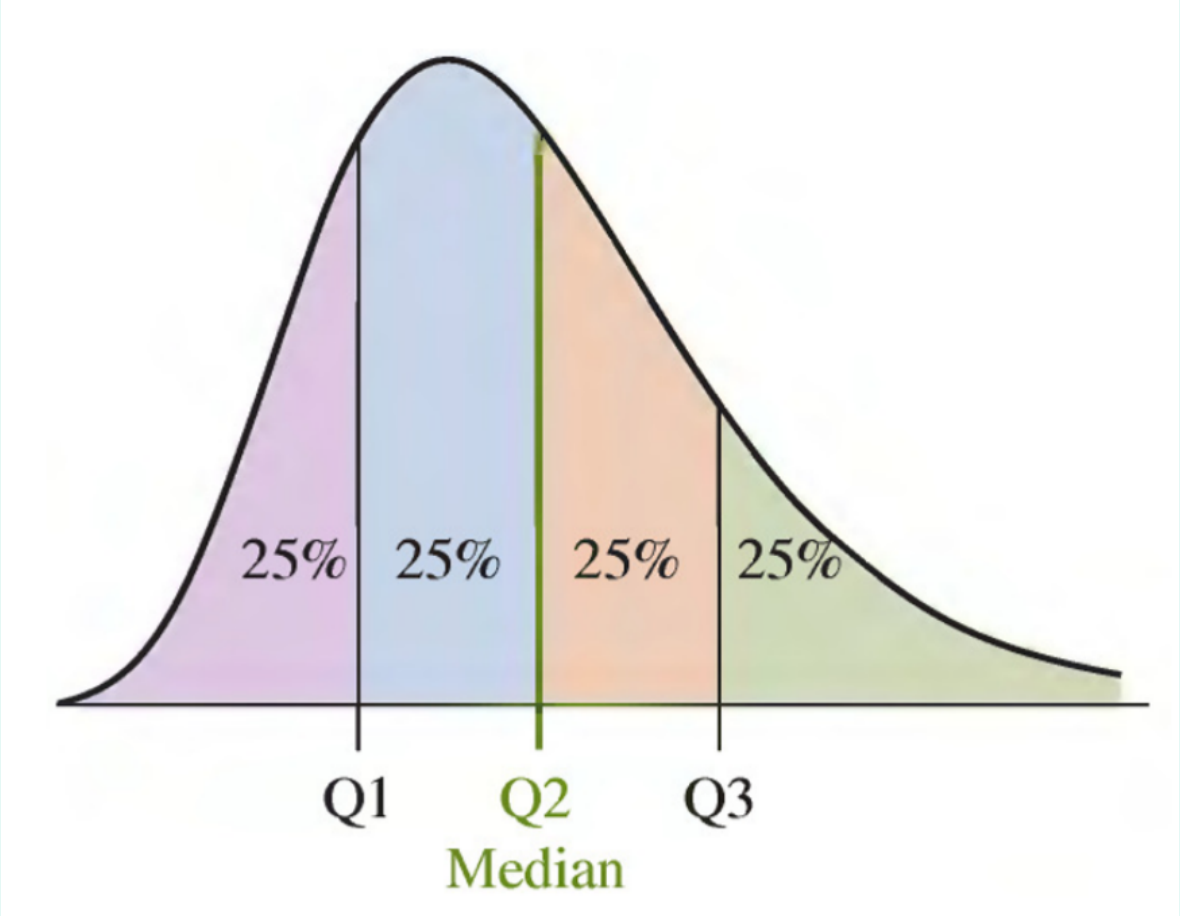
State	# of dentists
Ohio	57
Indiana	35
Illinois	61
Michigan	64
Wisconsin	90
Minnesota	78
Iowa	60
Missouri	53
North Dakota	55
South Dakota	57
Nebraska	71
Kansas	30

1. Compute the mean and the sample standard deviation.
2. Find the the 25th percentile (1st quartile), the 50th percentile (2nd quartile) and the 75th percentile (3rd quartile).
3. If in 2001 the number of dentists per 100,000 inhabitants reduces by 5% in all states, which are the new values for the mean and the sample standard deviation? And for the quartiles?
4. If the following transformation holds $y = 2x+3$, which are the new values for the mean and the sample standard deviation? And for the quartiles?
5. Compute the range and the interquartile range.
6. Draw the boxplot. What can we conclude by looking at this graph? Should any of the observed values be classified as potential outliers?

1. Mean and sample standard deviation:
    1. Mean: $\frac{57+35+61+64+90+78+60+53+55+57+71+30}{12} = 59.25$
    2. Sample standard deviation = $\sqrt{\frac{5.06+588.06+3.06+22.56+945.56+351.56+0.56+39.06+18.06+5.06+138.06+855.56}{11}}$ =
2. We got to split the observation in 4 subsets of equal length:
    1. First we sort the data: 30, 35, 53, 55, 57, 57, 60, 61, 64, 71, 78, 90.
    2. Then we do the thing: [30, 35, 53], [55, 57, 57], [60, 61, 64], [71, 78, 90]
    3. Then we compute the quartiles: $\frac{53+55}{2} = 54$, $\frac{57+60}{2} = 58.5$, $\frac{64+71}{2} = 67.5$

Quartiles recap

The Quartiles split the distribution into four parts that have the same number of observations:



You can find the quartiles by:

1. Ordering the set
2. Splitting the set in 4 subsets
3. Getting the mean between the extremes of the subsets

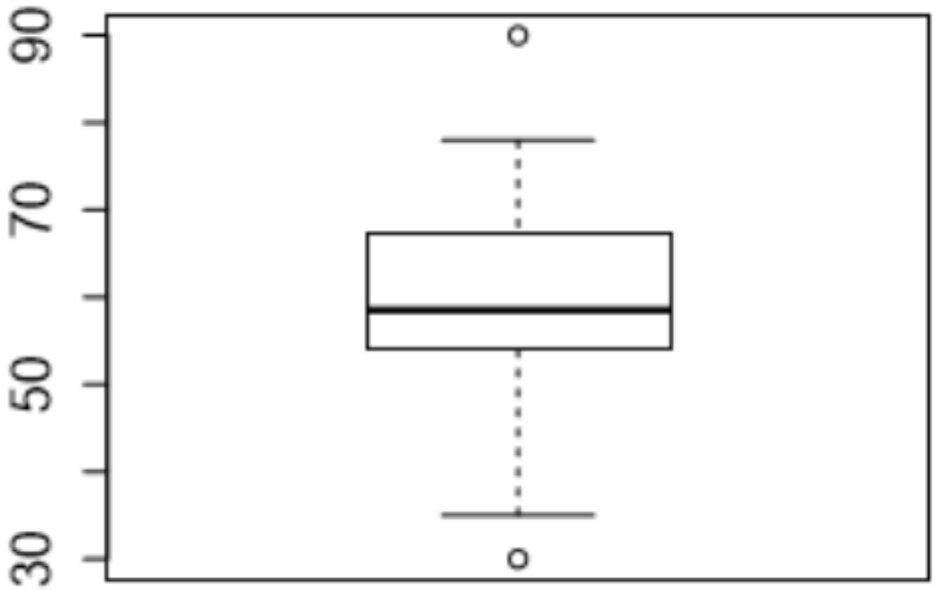
Example:

[30, 35, 53], [55, 57, 57], [60, 61, 64], [71, 78, 90]

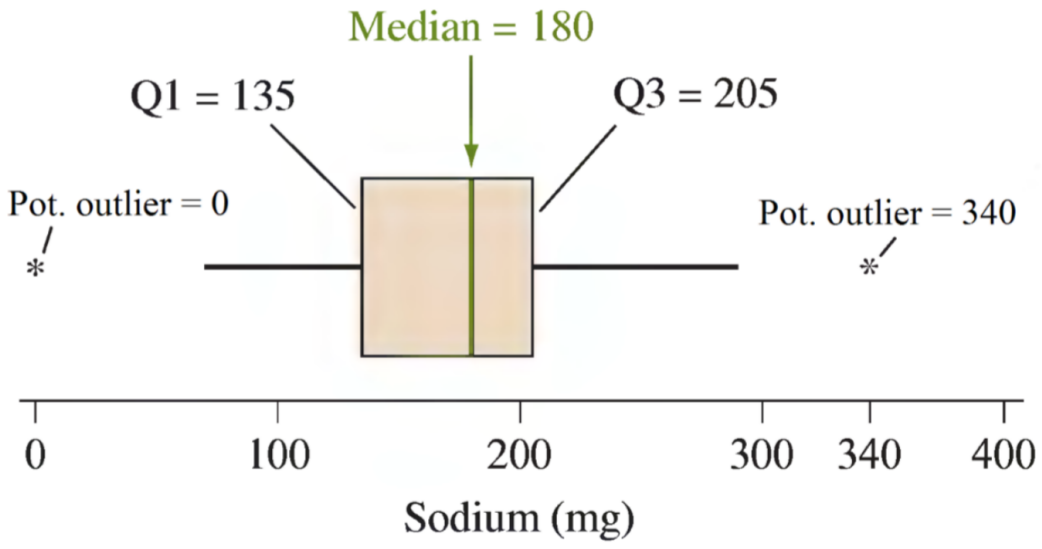
- Q1 = $\frac{53+55}{2} = 54$
- Q2 = $\frac{57+60}{2} = 58.5$
- Q3 = $\frac{64+71}{2} = 67.5$

3. The mean, the quartiles, and the standard deviation all reduce by 5%:
    1. New $\bar{x} = 0.95 * 59.25 = 56.29$,
    2. New $\sigma_s = 0.95 * 16.44 = 15.62$
    3. New $Q1 = 0.95 * 54 = 51.3$,
    4. New $Q2 = 0.95 * 58.5 = 55.58$,

    5. New $Q3 = 0.95 * 67.5 = 64.13$.
4. We do the same thing for some reason:
    1. New $\bar{x} = 2 * 59.25 + 3 = 121.5$,
    2. New $\sigma_s = 2 * 16.44 = 32.882$,
    3. New $Q1 = 2 * 54 + 3 = 111$,
    4. New $Q2 = 2 * 58.5 + 3 = 120$,
    5. New $Q3 = 2 * 67.5 + 3 = 138$.
5. Range and IRQ.
    1. Range is just the max - min value.
    2. IRQ is the length of the interval [Q1, Q3].
6. Box pot:



Example

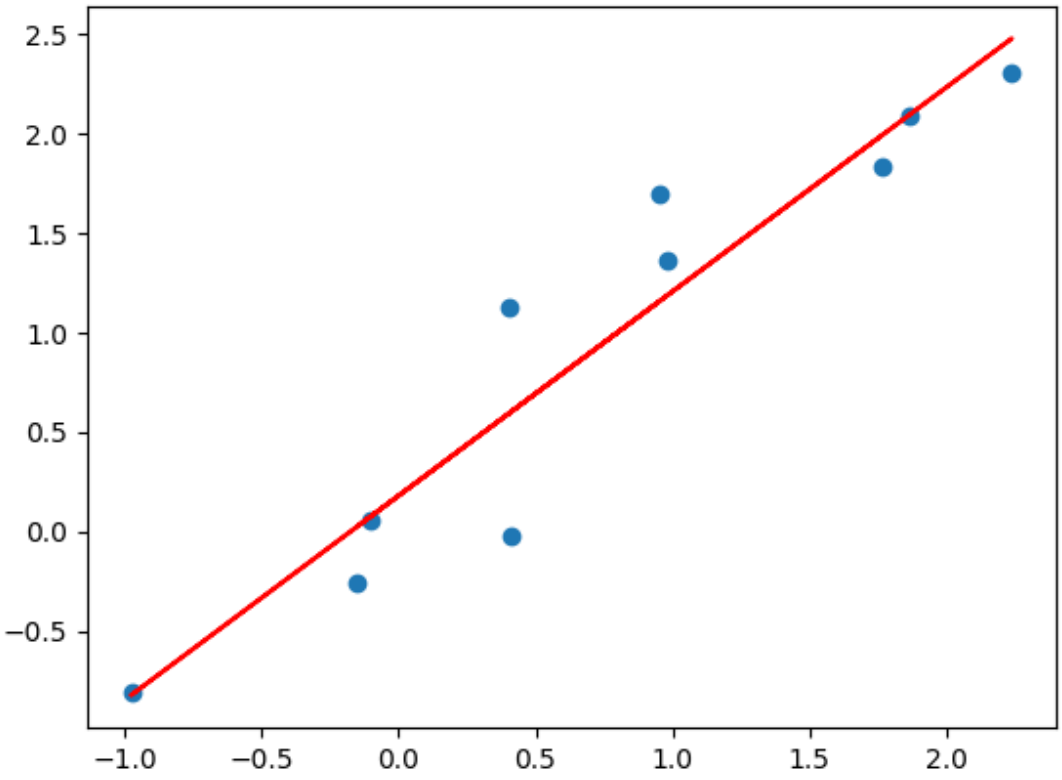


I can't do all exercises, from here i just categorize them

Recitation 3

Regression line

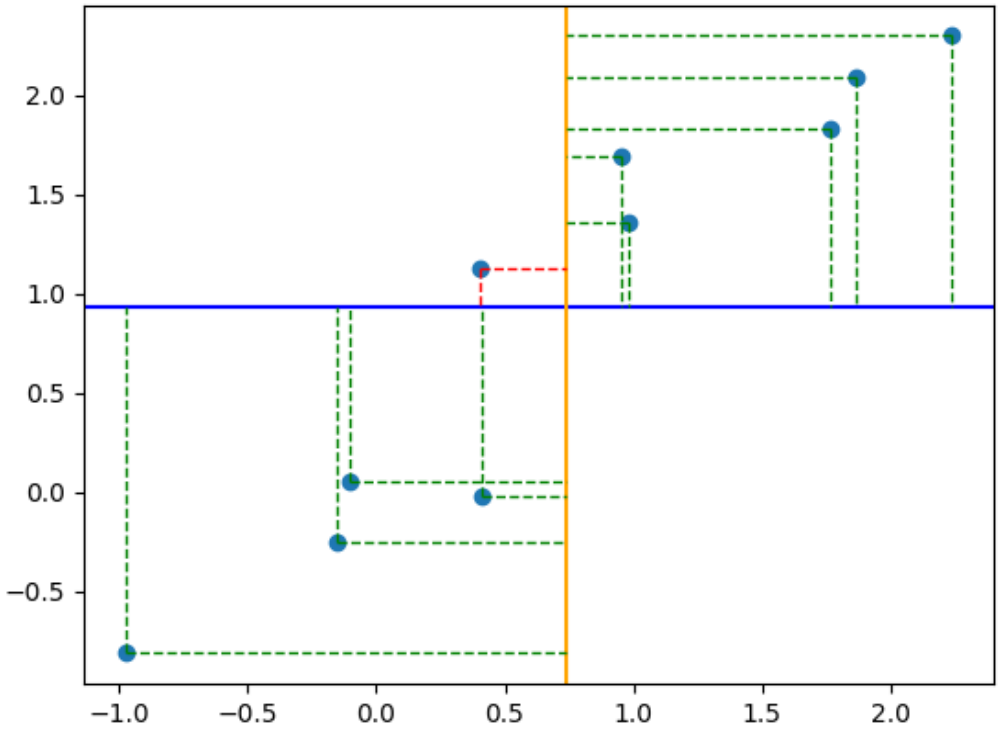
Imagine need to find that line:



We don't have to bruteforce our way to the best-fitting line.
We already know that a line can be described with the following formula:

$$y = mx + q$$

1. We find r (correlation coefficient):
    1. We find the covariance: $cov(x, y) = \sum(x - \bar{x})(y - \bar{y})$



2. Then:

$$r = \frac{cov(x,y)}{\sqrt{\sum(x-\bar{x})^2 \sum(y-\bar{y})^2}}$$

2. We find m:

$$m = r(\frac{\sigma_y}{\sigma_x})$$

3. We find q using the mean values, because we are sure that they are on the regression line:

$$y = mx + q$$

$$q = \bar{y} - m\bar{x}$$

> **⚠ Warning**
> If the exercise asks you to motivate why the regression line fits the data well, you say that $r^2 = tot\%$.
>
> $tot\%$ of the variability of X is explained by Y.

Residual

$$\text{e} = \text{prediction} - \text{ground truth}$$

Recitation 4

Sample space

It is the collection of all possible outcomes of an experiment.

A box contains four balls: one red, one blue, one yellow and one pink.

- **Consider an experiment that consists of drawing a ball from the box at random, replacing it, and drawing a second ball:**
  S = {RR, RB, RY, RP, BR, BB, BY, BP, YR, YB, YY, YP, PR, PB, PY, PP}
- **Let A be the event that the first ball drawn is Yellow. List all outcomes in A:**
  A = {YR, YB, YY, YP}
- **Let B the event that both balls have the same color. List all the outcomes in B:**
  B = {RR, BB, YY, PP}

Probabilities

We are in the same sample space as the examples above:

1. **Compute P(A) and P(Ac):**
   P(A) 1/4 = 0.25
   P(Ac) = 1 - 0.25 = 0.75
2. **Compute P(B) and P(Bc):**
   P(B) = 1/4 = 0.25
   P(Bc) = 1 - 0.25 = 0.75
3. **Compute P(A and B):**
   P(A and B) = P(A ∩ B) = 1/16 = 0.0625
4. **Compute P(A or B):**
   P(A or B) = P(A ∪ B) = 0.25 + 0.25 - 0.06 = 0.44
5. **Compute P(A | B):**
   P(A | B) = 0.0625/0.25 = 0.25

> **Quick recap on conditional probability:**
> When we are searching for the probability of an event A given that another event B has already happened, we can restrict the sample space to the event B.
>
> Now we search for the intersection of the two events in the sample space of B and we get this formula:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

Dependend and Independent Events

Two events are dependent if the happening of one of them changes the probability of the other one happening.

$$P(A \cap B) = P(A) \times P(B \mid A)$$

In this formula, the more B is independent from A, the more $P(B \mid A)$ approaches $P(B)$.

$$P(A \cap B) = P(A) \times P(B)$$

Wore Seat Belt	Survived (S)	Died (D)
Yes (Y)	410	5
No(N)	160	15

P(N) = 175 / 590 = 0.296
P(S) = 570 / 590 = 0.966
P(S|N) = 160 / 175 = 0.914

- **Compute the probability that an individual did not wear a seatbelt and survived:**
  P(N and S) = P(N) x P(S|N) = 0.296 x (160/175) = 0.27
  P(N and S) if independent = P(N) x P(S) = 0.296 x 0.966 = 0.286

  Since those two are not equal, the events are not independent.

Recitation 5

Probability distributions

We are sometimes asked to find the mean of a probability distribution. That is the Expectation:

$$E[X] = \mu = \sum p(x)\, x$$

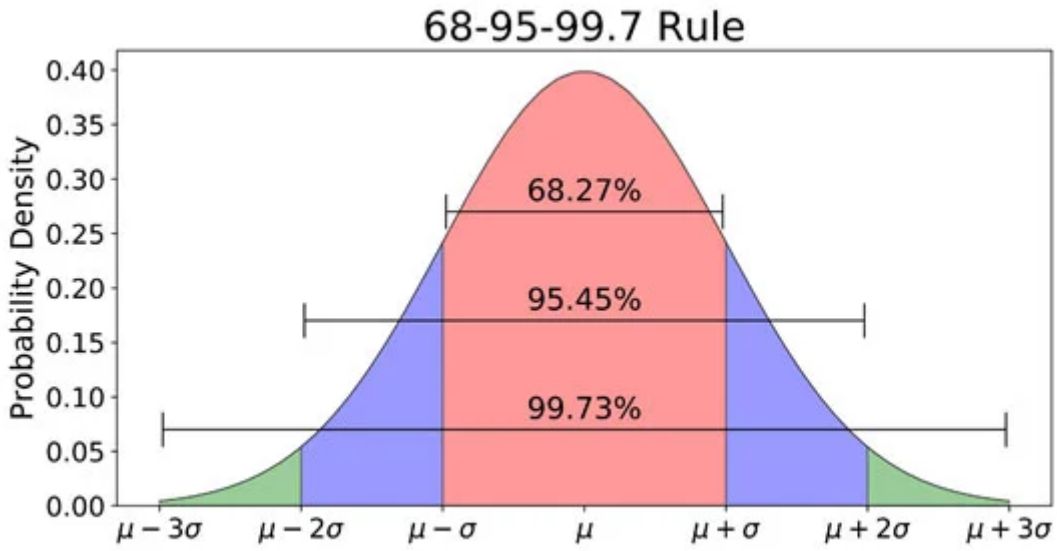
If we are asked to find the variance:

$$\sigma^2 = \sum (x - \mu)^2 P(x)$$

> **Tldr**
> Literally the expectation of the squared difference of the points from the mean.
>
> $$E[(X - \mu)^2]$$

> **Note**
> If you are asked to complete the distribution, remember that the y values must amount to 1.

Normal curve and Z-Score



Some exercises may ask you to calculate the probability in an interval of the Normal distribution.

I think you would do this with integrals, but i guess that integrating the normal distribution might not be easy?

Anyway, we have this exercise:

In a population the vehicle speed distribution is well approximated by a Normal curve with mean 50 and standard deviation 15.

- **Compute the probability that a randomly selected vehicle speed is greater than 73**

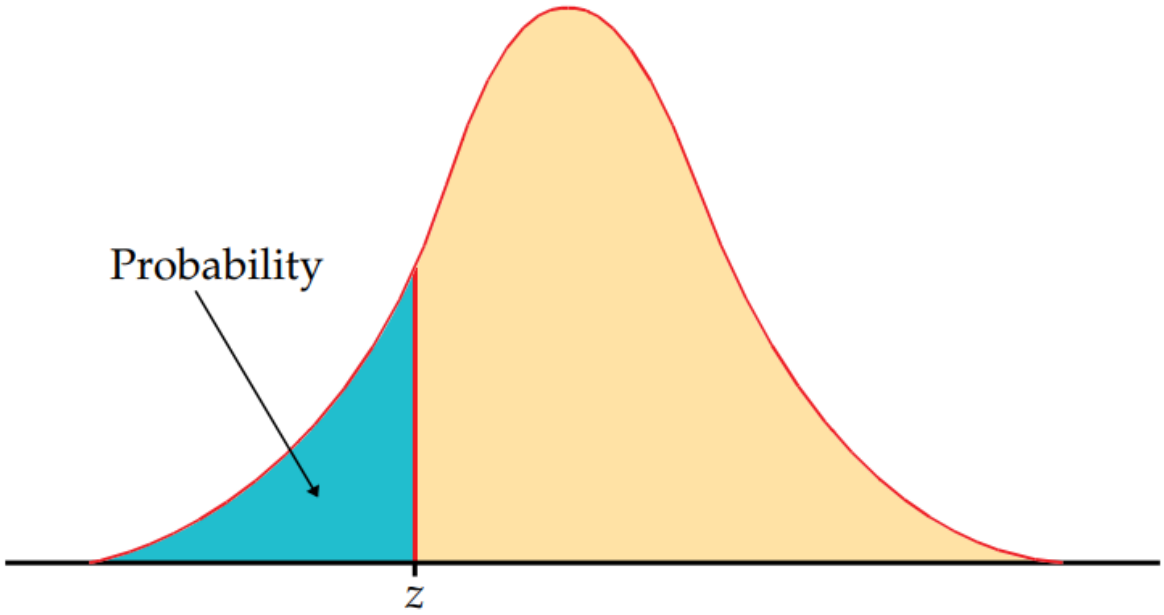
What is a z-score and what's its purpose:

Basically the z-score is how many standard deviations our value is away from the mean. The z-score is useful because it is standardized for every normal curve.

Basically if we get an exercise like the one above, where we would need to use an integral, we have a table of ready-to-go values, the z-table.

The z-table:

The z-table assigns to every z-score the area under the curve up to that z-score(the left of it).
The table is computed from the standard normal curve, but the z-score is standardized, so if our distribution follows a normal curve we can use the table.

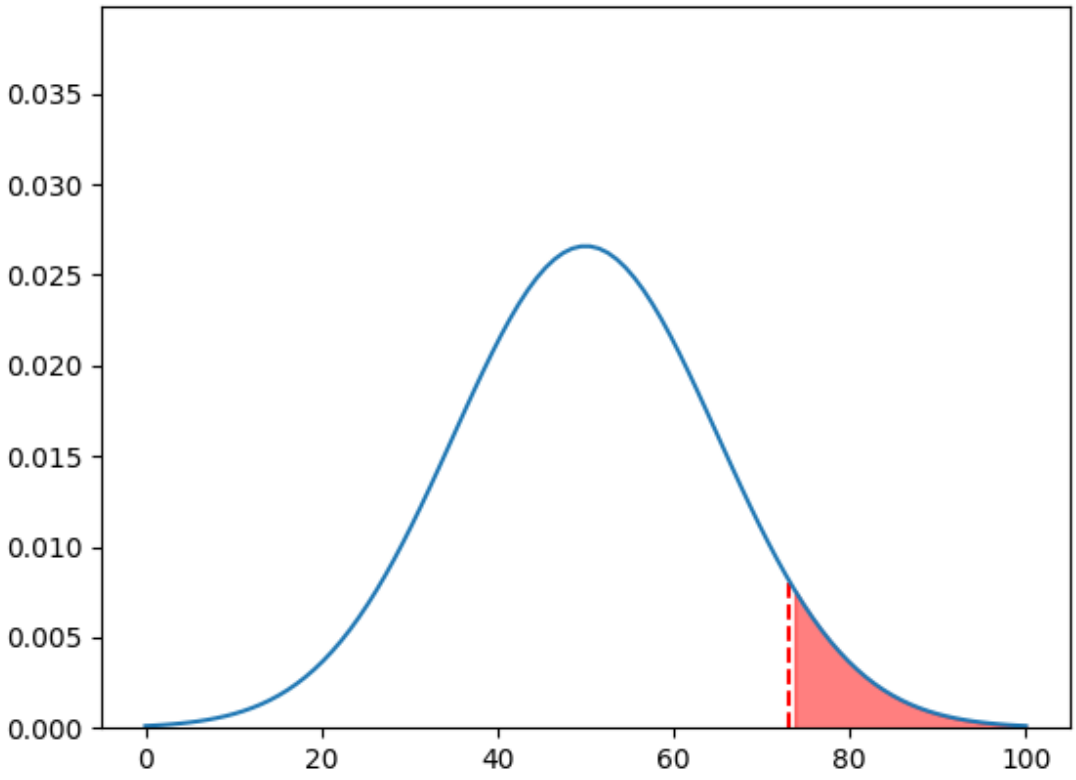


Getting back to the exercise:

Compute the probability that a randomly selected vehicle speed is greater than 73:

1. We compute the z-score:

$$z = \frac{x - \mu}{\sigma} = 1.53$$



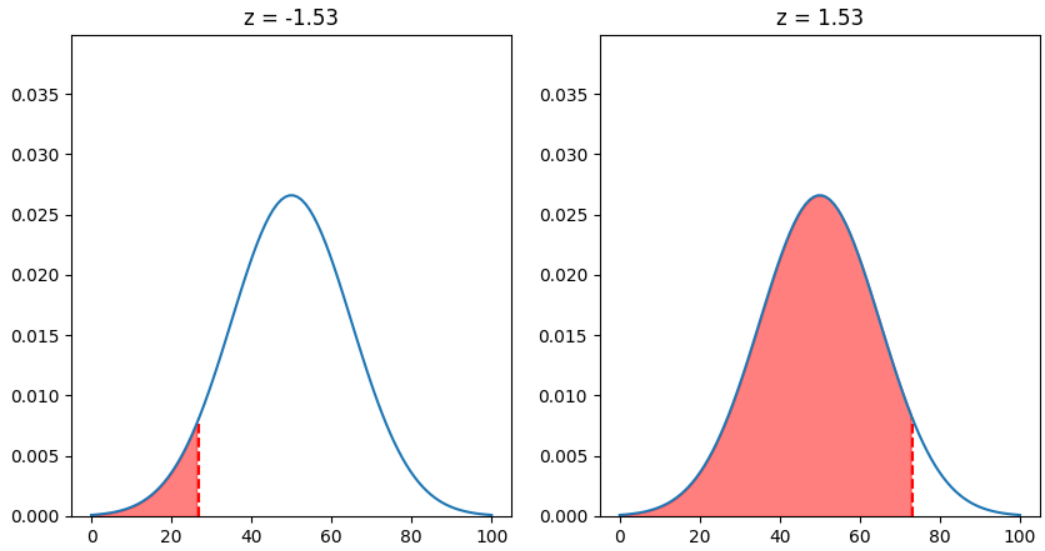
2. Now we've got to use the z-table to find the area corresponding to the z-score of 1.53:

TABLE A

Standard normal probabilities

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
−3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
−3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
−3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
−3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
−2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
−2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
−2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
−2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
−2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
−2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
−2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
−2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
−2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
−2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
−1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
−1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
−1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
−1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
−1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559

We use the entry -1.5 because if we used 1.53 we would get all the area to the left of 1.53, and we want the area to the right. We can use the negative z because the normal distribution is symmetric:



Our result is **0.063**.

> **Info**
> We could have also computed the complement of the area, instead of getting the inverse of the z-score.
>
> This is because the total area of the normal curve is 1.

Bernoulli distribution

A binomial distribution can be thought of as simply the probability of a SUCCESS or FAILURE outcome in an experiment or survey that is repeated multiple times.

Binomial distributions must meet the following three criteria:

1. **The number of observations or trials is fixed.**
   In other words, you can only figure out the probability of something happening if you do it a certain number of times. This is common sense:

```
- If you toss a coin once, your probability of getting a tails is 50%.
- If you toss a coin a 20 times, your probability of getting a tails is very, very close to 100%.
```

2. **Each observation or trial is independent.**
   In other words, none of your trials have an effect on the probability of the next trial.
3. The **probability of success** (tails, heads, fail or pass) is **exactly the same** from one trial to another.

The thing is imagine we flip a coin 5 times. There are $2^5 = 32$ equally likely outcomes. Now we want to know how many outcomes have 3 heads in them.
The formula to find out is:

$$P(x) = \frac{n!}{(n-x)!x!}p^x q^{n-x}$$

Where:

- n is the number of trials or the size of the sample.
- x is the number of successes, or in this case the number of heads.
- p is the probability of success in one trial.
- $q = 1 - p$, or the probability of failure in one trial.
- $P(x)$ is the probability of getting x successes(here 3 heads) in n trials with p probability of success per trial.

Mean of a binomial distribution is given by:

$$E[X] = np$$

Variance of a binomial distribution is given by:

$$var(X) = np(p-1)$$

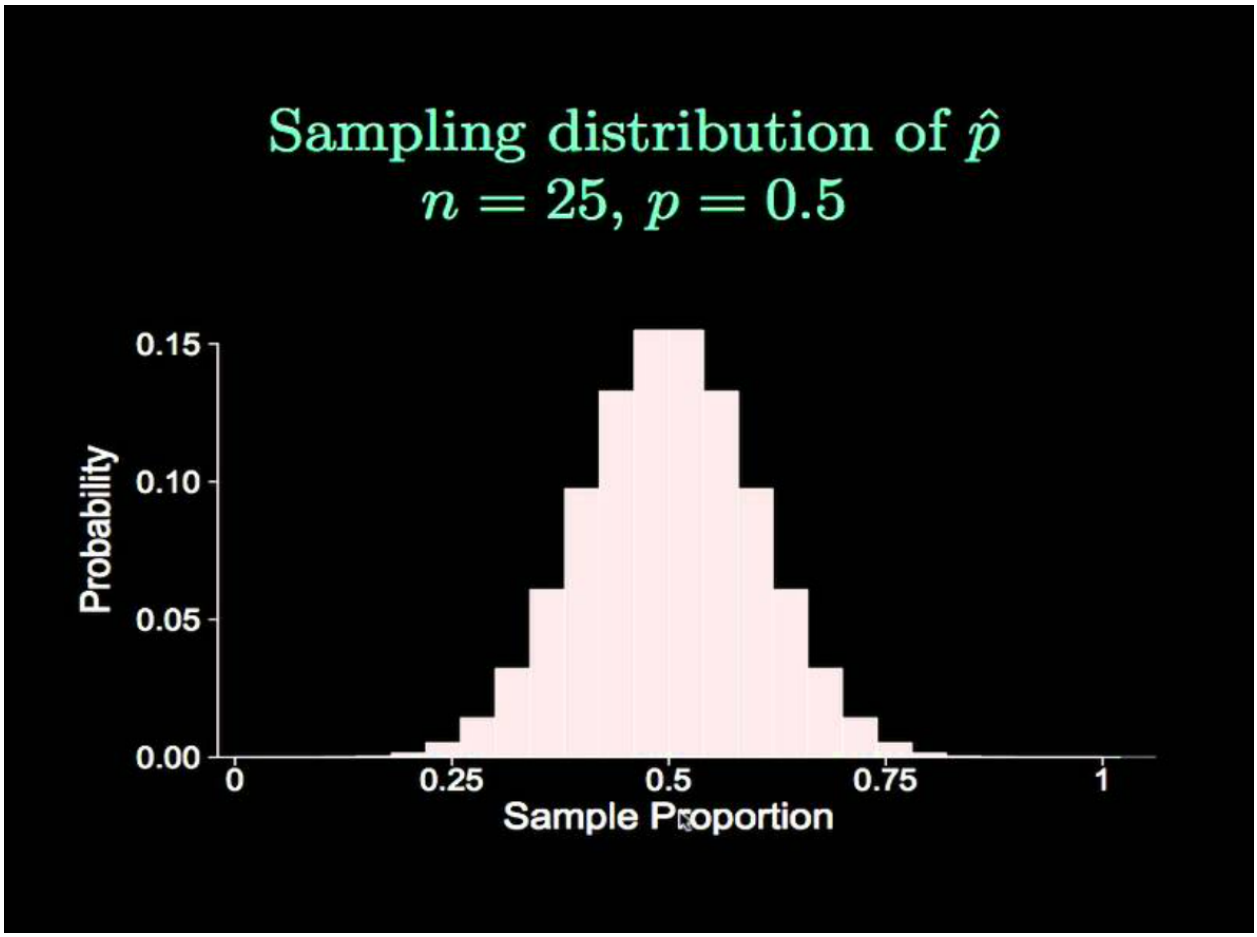
> **Hint**
> With big sample sizes this formula approximates to a normal distribution, so we can use z-scores to find areas under the curve.

Recitation 6

Sampling distributions

When people enter an apple store, $p = 0.25$ of them buys a product before leaving. This is the real proportion, it is the ground truth and we have it.

Imagine we sample the population and try to obtain p from the samples. Now P becomes uncertain, it is random variable $\hat{P}$:



According to the Central Limit Theorem, for large samples, the sample proportion is approximately normally distributed, with mean:

$$\mu_{\hat{P}} = p$$

and standard deviation of a proportion:

$$\sigma_{\hat{P}} = \sqrt{\frac{pq}{n}}$$

Where:

- p is the proportion/statistic of something.
- $q = 1 - p$
- $\mu_{\hat{P}}$ is the mean of the distribution of sampled proportions.
- $\sigma_{\hat{P}}$ is the standard deviation of the sampled proportions.

If we are investingating a mean(not a proportion) the formula for standard deviation is:

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

> ⚠ **Warning**
> Sometimes we want to compute the probability of successes being more than a certain number.

We know that we can get the area under a curve by using the z-scores, but this distribution only approximates a normal distribution when using a large n.

So when we have a small n we need to go sideways:

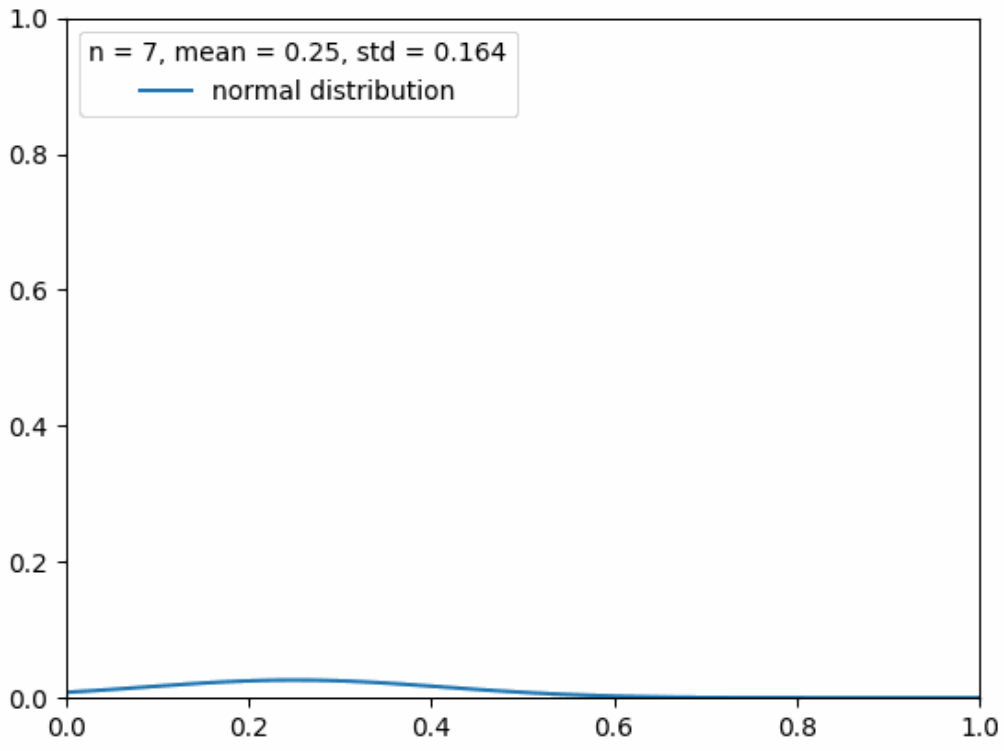
- If our configuration is also a binomial distribution, we can use that formula to compute every single discrete probability.

Example

For the population of individuals who own an iPhone, suppose p = 0.25 is the proportion that has a given app.

1. For a random sample of size n = 4, and the mean and the standard deviation of the sampling distribution of the sample proportion:

$$\mu_{\hat{P}} = 0.25, \quad \sigma_{\hat{P}} = \sqrt{\frac{0.25 \times 0.75}{4}} = 0.216$$



2. Find the probability that the proportion of having the app is at least 0.75 when n = 4.

   Here the sample size is too small, so we can't use the normal distribution stuff. 0.75 of 4 = 3, so we need the probability that at least 3 people have the app. We do that by summing the probabilities that 3 people have the app and 4 people have the app.

Since those probabilities are discrete and there are only 2 possible outcomes per trial, we can use the binomial distribution formula

Example

In the population, IQ scores are normally distributed with mean μ = 100 and variance σ 2 = 15. Suppose to draw a random samples of 25 individuals from the population and measure the IQ score

1. Compute the probability of observing a sample mean between 98 and 102 when drawing a sample of 25 individuals:

The standard deviation of the sample mean for a sample of size 25. Is given by:

$$\sigma_s = \frac{\sigma}{\sqrt{n}} = \frac{\sqrt{15}}{\sqrt{25}} = 0.774$$

The z-scores for 98 and 102 are:

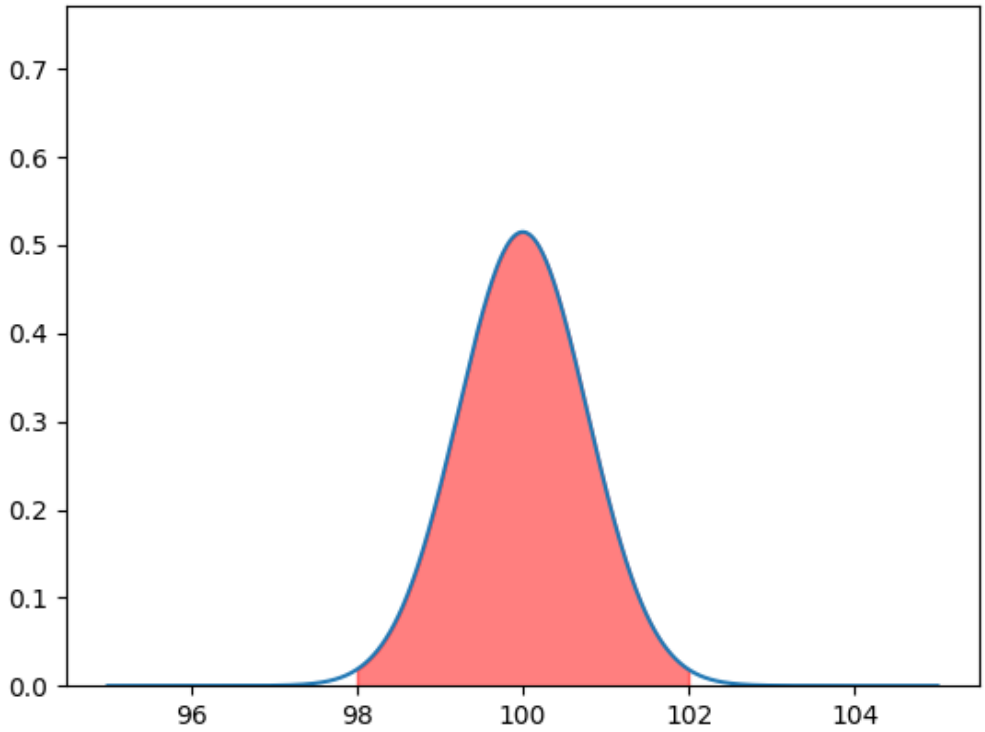
$$z_1 = \frac{98 - 100}{0.774} = -2.582, \quad z_2 = \frac{102 - 100}{0.774} = 2.582$$

The areas given by the z-tables for the z-scores are:

$$A_{z1} = 0.0049, \quad A_{z2} = 0.9950$$

The area between the two z-scores is given by:

$$A_{z2-z1} = A_{z2} - A_{z1} = 0.9950 - 0.0049 = 0.9901$$



Recitation 7

Standard Error

It's the average error of the estimation from the samples.

For the sample mean it is:

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

For the sample proportion it is:

$$e = \sqrt{\frac{pq}{n}}$$

ⓘ What is the Difference between Standard Error and Standard Deviation?
Standard error and standard deviation are both measures of variability, but standard deviation is a descriptive statistic that can be calculated from sample data, while standard error is an inferential statistic that can only be estimated.

Confidence interval

Confidence Level

The confidence level is the overall capture rate if the method is used many times. The sample mean will vary from sample to sample, but the method estimate ± margin of error is used to get an interval based on each sample.
C% of these intervals capture the unknown population mean μ.
In other words, **the actual mean will be located within the interval C% of the time.**

The population mean for a certain variable is estimated by computing a confidence interval for that mean.

The formula for the confidence interval is:

$$\text{Confidence interval} = \text{sample mean} \pm \text{margin of error}$$

Example:
Sample mean=80
Margin of error=3.92
Confidence level (CI): 95%

CI (95%)= 80± 3.92
Upper limit=80+3.92=83.92
Lower limit=80-3.92=76.08

CI(95%)=(76.08; 83.92)

*Researchers can say, with 95% confidence that the population mean on the motivation scale is between 76.08 and 83.92.

In order to find the confidence interval, we must find the margin of error first:

$$\text{Margin of error} = z^* \cdot \text{Standard Error}$$

Formula Explanation
For each C% there is a specific z-score, that gives you the bounds of the interval.

The margin of error is just the un-normalized bound, because we are converting the z-score to a real value.

Where to find values for z^*?
You can find the values for z^* in the c-table or z-table.

Ex. **When a General Social Survey asked 1326 subjects, "Do you believe in science?", the proportion who answered yes was 0.82.**

Construct the 95% confidence interval.

The sample proportion is equal to 0.82, now we just need the standard error in order to calculate the margin of error.

$$\text{Standard error} = \sqrt{\frac{0.82 \times 0.18}{1326}} = 0.011$$

Why this formula?

This is the standard error formula for the sample proportion:

$$e = \sqrt{\frac{pq}{n}}$$

It's different from the one we use for the sample mean:

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

$$\text{Margin of error} = 1.96 \cdot 0.011 = 0.0215$$

So the confidence interval would be

$$CI_{95\%} = 0.82 \pm 0.0215 = [0,7985, 0,8415]$$

How to interpret

We are 95% confidence that between 79.8% and 84.2% of people believe in science.

Hint

If sample size increases, the margin of error decreases, and thus the CI becomes narrower.

Hint

Describe the effect of standard deviation, sample size and α on the confidence interval.

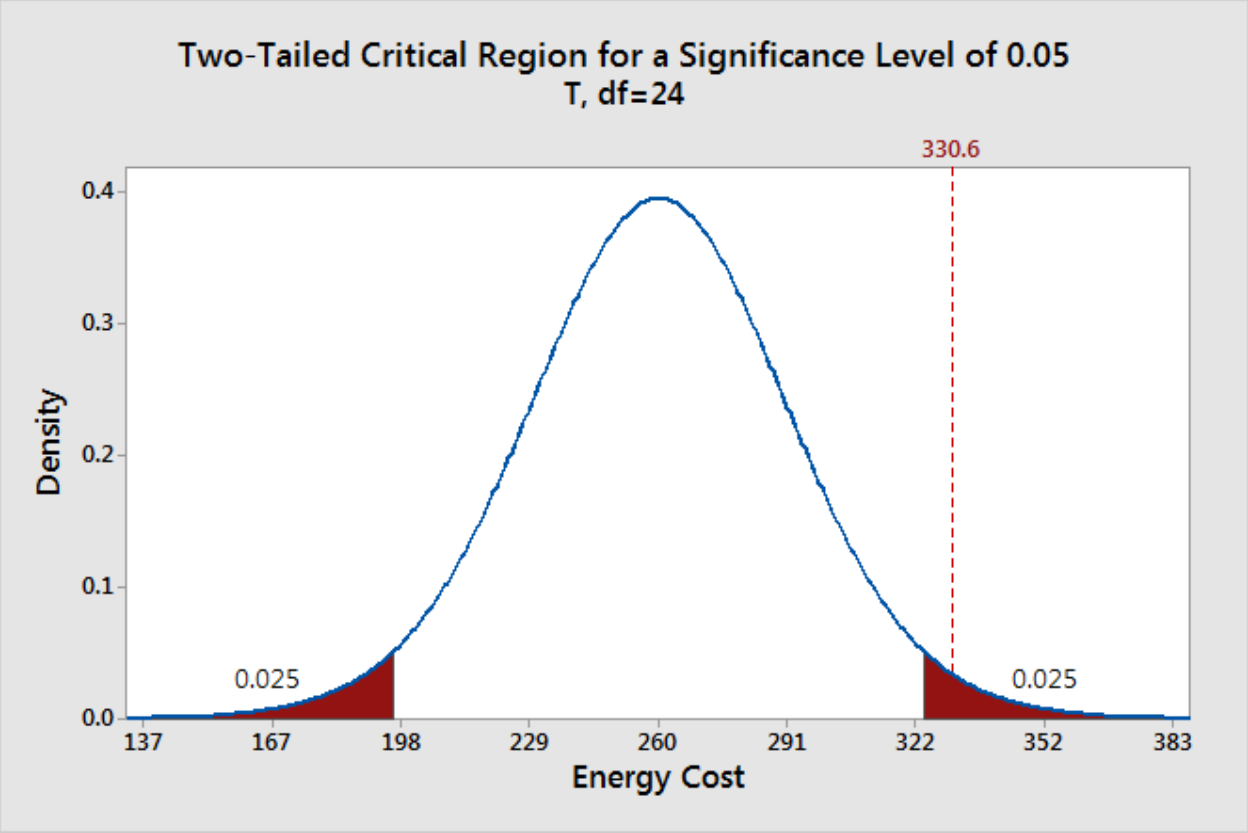
- lower standard deviation → lower margin of error → narrower CI
- higher standard deviation → higher margin of error → wider CI
- lower sample size → higher margin of error → wider CI

- higher sample size → lower margin of error → narrower CI
- lower α → higher level of confidence 1-α →higher margin of error → wider CI
- higher α → lower level of confidence 1-α → lower margin of error → narrower CI

Significance Level

The significance level is a threshold that determines whether a study result can be considered statistically significant after performing the statistical tests.

A α of 0.05 indicates a 5% risk of concluding that a difference exists between the population mean and the sample mean, when there is no actual difference(so the probability of getting a bad sample).



We expect to obtain a sample mean that falls in the critical region 5% of the time.

Multiplier for significance level

I think it is the z-value for the bounds of the critical regions.
If $\alpha = 0.05$, then a single area measures $\frac{\alpha}{2} = 0.025$.

So we need to find the z-value for the area 0.025 and that will be our "multiplier".

t-distribution

It's like a normal curve but wider.
You must use the t-distribution table(instead of z-table) when the population standard deviation is not known and the sample size is small (n<30).

General Correct Rule: If σ is not known, then using t-distribution is correct. If σ is known, then using the normal distribution is correct.

> **Danger**
> Do exercise 2 and the last ones. We did not exercises on t-distribution.

Recitation 8

Significance Test

I will explain using an exercise:

In a sample of 402 Tor Vergata first-year students, 174 are enrolled into Statistics course.

Is the proportion of students enrolled into Statistics course in the population of all Tor Vergata first-year students different from 0.50 at the significance level α = 0.05?

1. Assumptions

The distribution approximates to a normal distribution because of the large sample size.

2. Hypothesis

We are given an hypothesis:

$$H_0 : p = 0.5, \quad H_1 : p \neq 0.5$$

Where:

- H_0 is the actual hypothesis, or null hypothesis.
- H_1 is the alternative hypothesis.

In a significance test, the null hypothesis is presumed to be true unless the data give strong evidence against it.

3. Test statistic

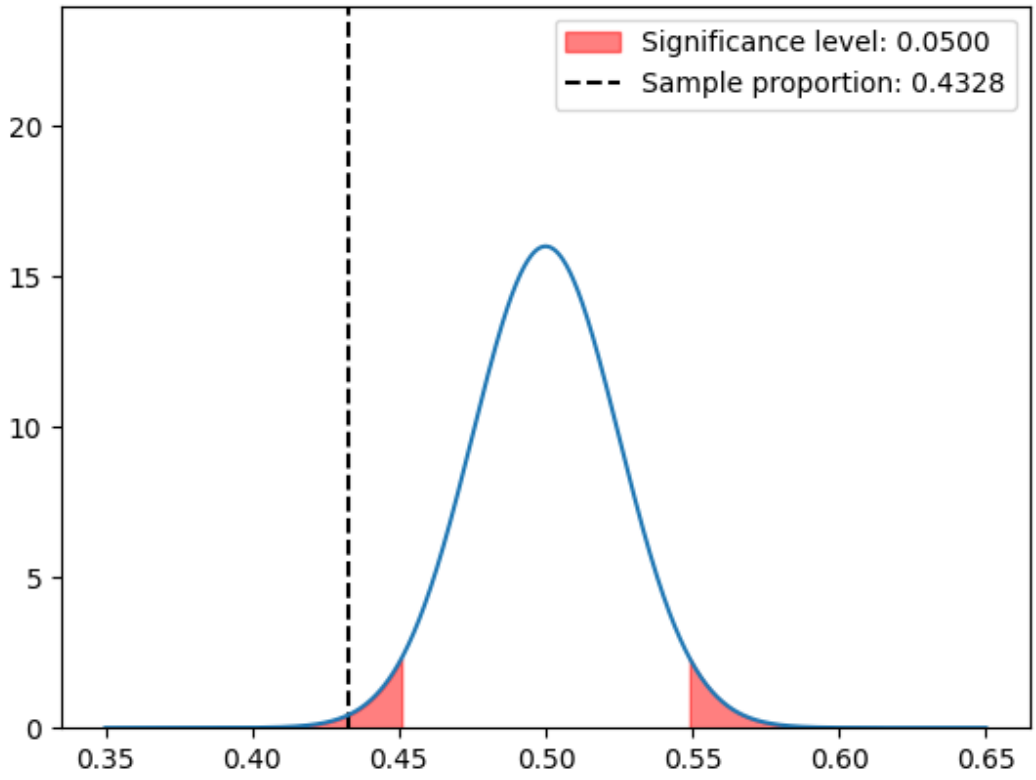
A test statistic measures how far the point estimate falls from the parameter value given in the null hypothesis. The result is the number of standard errors between the two.

First, we construct the normal curve considering the hypothesis:

$$n = 402, \quad p = 0.5, \quad \sigma = \sqrt{\frac{pq}{402}} = 0.0249$$

Then we take the sample:

$$\hat{p} = \frac{174}{402} = 0.4328$$



We can already see from the plot that this sample proportion really doesn't agree with our hypothesis.
Mathematically, to disprove the hypothesis we need to check if the sample mean/proportion lands beyond the significance level threshold.

In order to do just that, we need the z-score for the sample proportion, also called the test statistic:

$$z = \frac{0.4328 - 0.5}{\sigma} = -2.6947$$

Now we compute the areas under the curves to determine whether the sample proportion exceedes the α threshold or not.

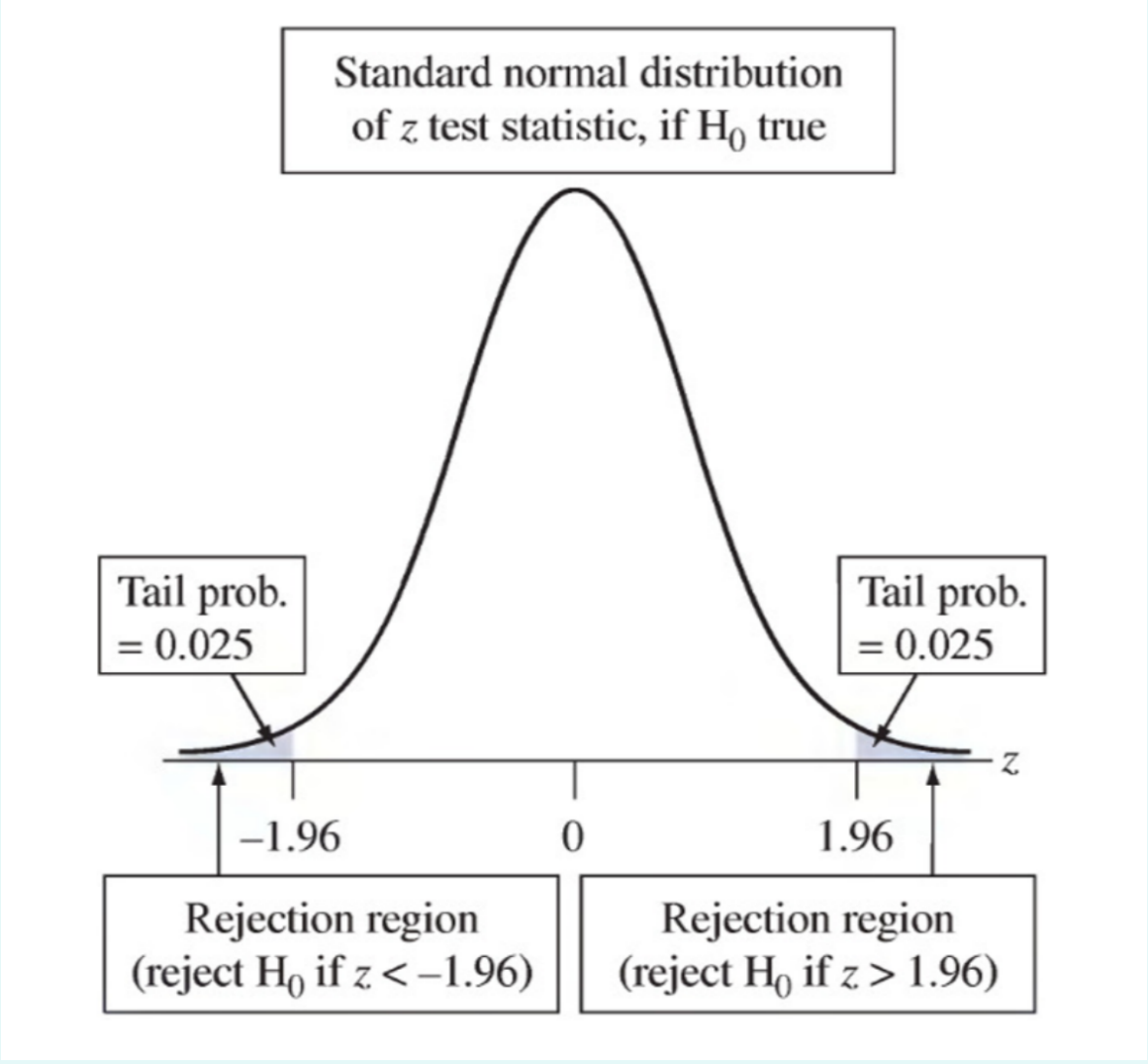
4. p-value

This is just the area under the curve at the left of the test statistic value. We can later compare it with the significance level to see if we really are out of bounds.

$$\text{p-value} = P(Z < \text{test statistic}) \times 2 = P(Z < -2.6947) \times 2 = 0.0035 \times 2 = 0.007($$

Why x2?

Because the significance level α is the area under both tails of the distribution. So we can either use $\frac{\alpha}{2}$ or $\text{p-value} \times 2$ when comparing.



5. Conclusion

Once we have the p-value, we proceed to either accept or reject the hypothesis by comparing the two area

- If $\text{p-value} < \alpha$, we reject H_0
- If $\text{p-value} \geq \alpha$, we accept H_0

In this case $0.0070 < 0.05$, so we reject the hypothesis.

Recitation 9

Simple regression equation

A regression equation describes how the mean value of a Y-variable relates to specific values of the X-variable:

$$E[y_i] = \beta_0 + \beta_1 \cdot x_i$$

or

$$Y = \beta_0 + \beta_1 \cdot X$$

SSE

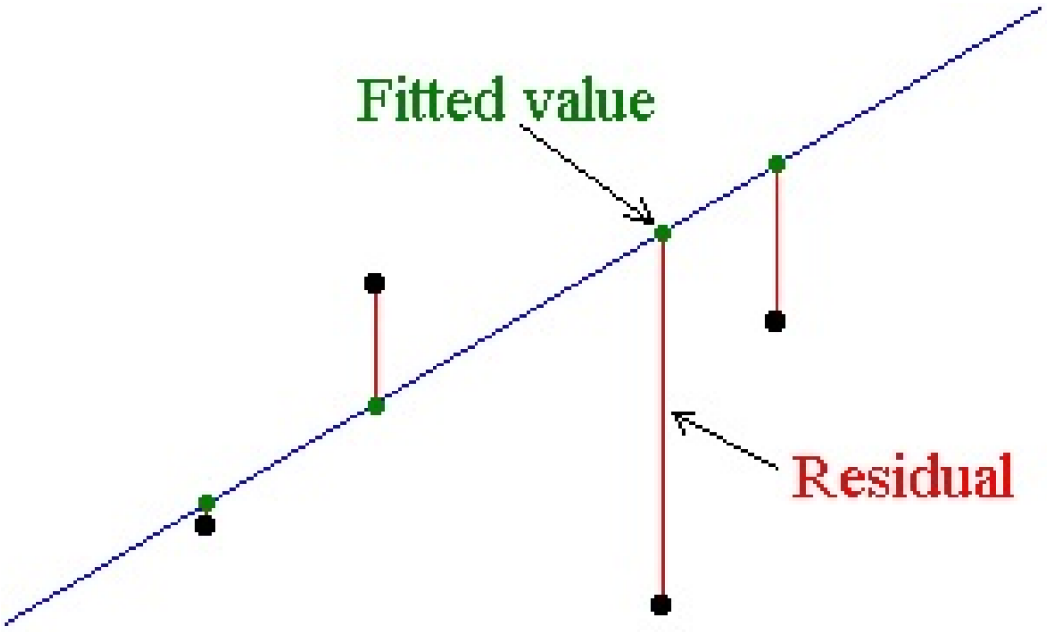
SSE stands for "Sum of Squared Errors", it is calculated as follows:

$$\text{SSE} = \sum_{i=0}^{N} (y_i - \hat{y}_i)^2$$

Where:

- $\hat{y}_i$ is the predicted value for x_i
- y is the ground truth for x_i

We are basically computing a sum of the residuals, once we have our prediction:



How to find b_0 and b_1

If we want to find the two parameters, we got to minimize the SSE.

We do that by computing b_0 and b_1 as:

$$b_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

$$b_0 = \bar{y} - b_1 \cdot \bar{x}$$

They are called the "least squares estimates" of β_0 and β_1.

> ⚠ **Warning**
> We use the notation β_0 and β_1 when we are talking about population/ground truth parameters.
>
> We use the notation b_0 and b_1 when we are talking about fitted/predicted parameters.

MSE

If we are trying to estimate how good our model is, we can use the "Mean Squared Error":

$$\text{MSE} = \frac{\text{SSE}}{n-p}$$

Where:

- n is ?
- p is the number of parameters, in this case 2.

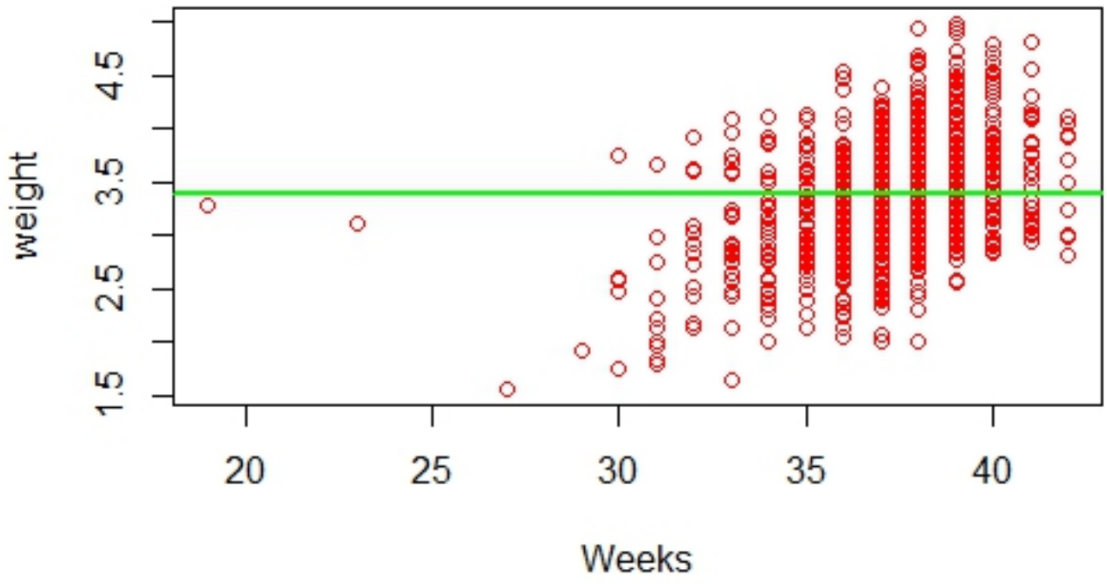
> **Hint**
> The MSE is the sample variance of the errors and estimates σ^2:
>
> $$E\{\text{MSE}\} = \sigma^2$$
>
> Of course, then the standard deviation of errors is computed as:
>
> $$s = \sqrt{\text{MSE}}$$
>
> Which is the sample standard deviation of the errors (residuals) from the regression line.
> It can be interpreted as the average deviation of individuals from the sample regression line.

SST

It is the total sum of squares:

$$\text{SST} = \sum_{i=0}^{N} (y_i - \bar{y})^2$$

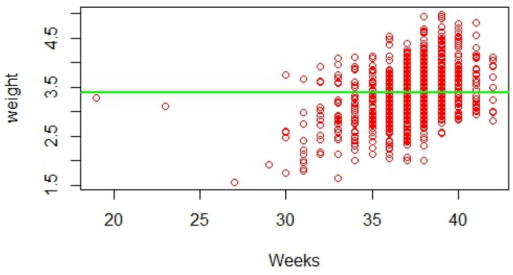
It is basically the sum of all the squared deviations from the mean.



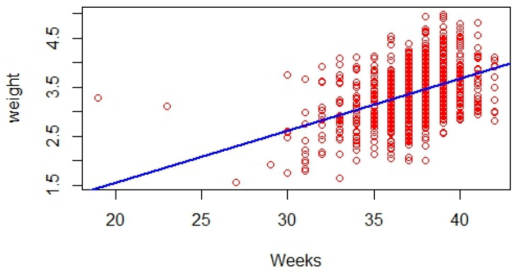
R^2

$\sqrt{R^2} = r \leftarrow$ correlation coefficient.

$$R^2 = \frac{\text{SST} - \text{SSE}}{\text{SST}}$$



green line:$\bar{y} = 3.385$,
$SSTO = 312.0133$



blue line:$y = -0.56 + 0.11x$,
$SSE = 254.2687$

> **Hint**
> It is interpreted as the fraction of variation in y that is explained by the fitted regression equation. It is often converted to a percentage.

Formulas

General formulas

Sample standard deviation

$$\sigma_s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{N-1}}$$

Simple regression

Covariance:

$$\operatorname{cov}(x,y) = \sum(x-\bar{x})(y-\bar{y})$$

Correlation coefficient

$$r = \frac{\operatorname{cov}(x,y)}{\sqrt{\sum(x-\bar{x})^2 \sum(y-\bar{y})^2}}$$

Angular coefficient

$$m = r(\frac{\sigma_y}{\sigma_x})$$

Intercept

$$q = \bar{y} - m\bar{x}$$

Normal distribution

z-score

$$z = \frac{x - \mu}{\sigma}$$

Test statistic(s-score with $\sigma = \mathrm{SE}$)

$$z = \frac{x - \mu}{\mathrm{SE}}$$

Bernoulli distribution

Probability at x

$$P(x) = \frac{n!}{(n-x)!x!}p^x q^{n-x}$$

Mean

$$\mu = np$$

Variance

$$\text{var}(X) = np(p-1)$$

Sampling distributions

Standard deviation of sample proportion

$$\sigma_{\hat{P}} = \sqrt{\frac{pq}{n}}$$

Standard error of sample proportion

$$e = \sqrt{\frac{pq}{n}}$$

Standard deviation of sample mean

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

Standard error of sample mean

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

Confidence intervals

Confidence interval

$$\text{Confidence interval} = \text{sample mean} \pm \text{margin of error}$$

Margin of error

$$\text{Margin of error} = z^* \cdot \text{Standard Error}$$

Significance level

Multiplier: Just the z-score of the bound of the significance level

Lil more aggressive Regression

SSE

$$\text{SSE} = \sum_{i=0}^{N} (y_i - \hat{y}_i)^2$$

SST

$$\mathrm{SST} = \sum_{i=0}^{N} (y_i - \bar{y})^2$$

R^2

$$R^2 = \frac{\mathrm{SST} - \mathrm{SSE}}{\mathrm{SST}}$$