MAT 2377 Summary Sheet

1 Chapter 1: Probabilities

The **sample space** is the set of all possible outcomes.

An **event** is a collection of outcomes in the sample space. Usually this is what we are looking to work with.

We can count items using the k stage procedure.

If we have k stages, each with n_1, n_2, n_3, ... possibilities, then the total number of possiblilites is just $n_1 \cdot n_2 \cdot n_3 \cdot ... \cdot n_k$.

1.1 Ordered Samples

If we have an ordered sample, then we see that picking $1, 2, 3$ is different than picking in a different order $1, 3, 2$.

If we draw r items from a bag of n items:

- If we replace each item after drawing, we have: $n \cdot n \cdot n \cdot ... = n^r$ possibilities
- If we do NOT replace the items, we have: $n \cdot (n-1) \cdot (n-2) \cdot ... \cdot (n-r) = \frac{n!}{(n-r)!} = {}_nP_r$

1.2 Unordered Samples

This is when the **order of the samples does not matter**, so $1, 2, 3$ would be the same as $1, 3, 2$.

We can see the number of unordered samples possible with r draws in a sample space of size n using:

$$\frac{n!}{(n-r)!r!} = {}_nC_r$$

Ex. 20 Items are sampled from a factory. The probability of an item being defective is 0.1. What is the probability of **exactly** 1 defective item?

To do this, we know that for there to be exactly 1 defective item, there must be 19 working items.

But there are more than one way that this 1 defective item can be placed. It can be the first item we test, the second item we test, and so on. We do this by $\binom{20}{1}$.

$$P(1 \text{ Defective Item}) = \binom{20}{1} 0.1^1 (1-0.1)^{19} = 0.27$$

If we changed this do exactly 2 items, the formula would change to:

$$P(\text{2 Defective Items}) = \binom{20}{2}0.1^2(1-0.1)^{18} = 0.285$$

1.3 Probabilities

The probability of an event A with N total outcomes and a favourable outcomes is just:

$$P(A) = \frac{a}{N}$$

We can **add probabilities** using the following formulas:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$
$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$$

Any 2 events that satisfy the following expression are called **independant.**

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

Ex. The probability of event A is 0.95, and of event B is 0.98. The probability of them both is 0.94. What is the probability of neither of them occurring?

We have:

$$P(A) = 0.95 \qquad P(B) = 0.98 \qquad P(A \cap B) = 0.94$$

We want:

$$P(A \cup B)\prime = 1 - P(A \cup B) = 1 - (P(A) - P(B) + P(A \cap B)) = 0.01$$

1.4 Conditional Probability

We say that the probability of event B **given that event** A **has already happened** is:

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

1.5 Law of Total Probability

This basically works off of the fact that **all probabilities must add up to 1**.

This is the specific case to 2 events A and B:

$$P(B) = P(B|A)P(A) + P(B|\overline{A})P(\overline{A})$$

This uses the fact that A and $\overline{A}$ are mutually exclusive, and exaustive (covers all of S).

So in general, if we have $A_1, A_2, ..., A_k$ and $A_1, A_2, ..., A_k$ are mutually exclusive and exaustive, then we say:

$$P(B) = P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + ... + P(B|A_k)P(A_k)$$

1.6 Bayes Theorum

This is a way to get the opposite conditional probability to what we have.

If we have $P(A|B)$, among a couple other things, we can obtain $P(B|A)$ with:

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

Ex. We have a disease. A test for the disease was given to 450 patients who are known to have this disease. 436 were positive. It was also given to 1000 people without the disease where 10 tests were positive. The known rate of this disease is 15%. What is the probability of someone having the disease given the test is positive?

I will let D represent Having the Disease, and P represent the Test being Positive.

We have the following information from the question:

$$P(P|D) = 436/450 \qquad P(P|D\prime) = 10/1000 \qquad P(D) = 0.15 \qquad P(D|P) = ?$$

We can use Bayes theorum to find $P(D|P)$

$$P(D|P) = \frac{P(P|D)P(D)}{P(P)}$$

To find the denominator, we can use the law of total probability.

$$P(P) = P(P|D)P(D) + P(P|D\prime)P(D\prime) = \frac{436}{450}0.15 + \frac{10}{1000}(1 - 0.15) = 0.1538$$

And now I can substitute back into Bayes theorum to solve:

$$P(D|P) = \frac{P(P|D)P(D)}{P(P)} = \frac{436/450 \cdot 0.15}{0.1538} = 0.945$$

2 Chapter 2: Discrete Random Variables

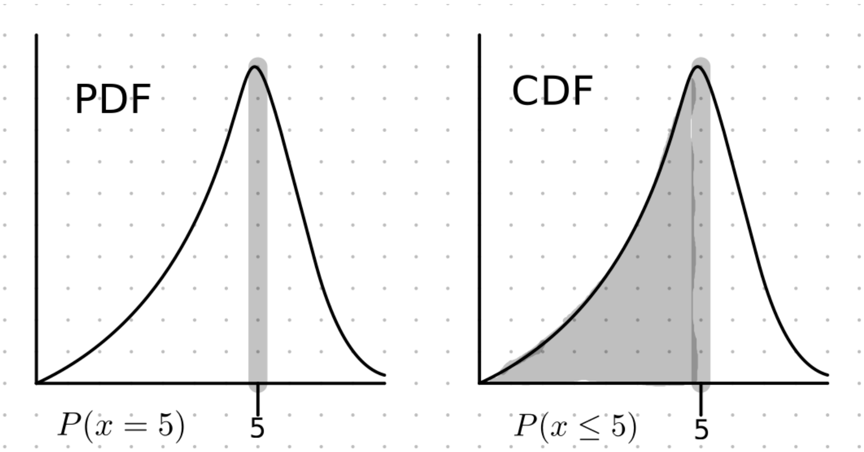
A random variable is a variable (typically a capital letter) that associates a **number to every outcome of an experiment**.

We have 2 functions:

- Probability Distribution Function (PDF) ($f(x)$)
- Cumulative Distribution Function (CDF) ($F(x)$)

The PDF specifies the probability of getting **this specific value**.

The CDF specifies the probability of getting **anything below this** specific value.



Ex. We have a standard fair D6.

We know that the PDF is $\frac{1}{6}$ for each value.

The CDF is a bit more complex in that it is $\frac{1}{6}$ for 1, $\frac{2}{6}$ for 2, up to $\frac{6}{6} = 1$ for 6.

$$P(X = 4) = \frac{1}{6} \qquad \text{Using PDF}$$

$$P(X \leq 4) = \frac{4}{6} \qquad \text{Using CDF}$$

$$P(3 \leq X \leq 5) = P(X \leq 5) - P(X \leq 3) = \frac{5}{6} - \frac{3}{6} + \frac{1}{6} \qquad \text{Using CDF}$$

For the third example, we need to add the extra $\frac{1}{6}$ since we are doing from 3 to 5 inclusively. So 3 or 4 or 5. Taking away $X \leq 3$ also takes away 3, so we need to add it back.

Note that the sum of all PDFs is always 1. The highest CDF is also 1.

2.1 Expectation

The expectation of a random variable is basically the **mean** of the variable. We say that this is:

$$\mathbb{E}(u(x)) = \sum u(x) P(X = x)$$

Ex. The expectation of a fair D6 is:

$$\mathbb{E}(X) = \sum xP(X = x) = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + ... + 6 \cdot \frac{1}{6} = 3.5$$

2.2 Variance

While the mean shows the average of the data, the variance shows **how close most of the data is to this average**. This can be calculated using the expectation formula:

$$\text{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$$

We define the standard deviation as the square root of the variance.

$$\text{SD}(X) = \sqrt{\text{Var}(X)}$$

Ex. Find the variance for the fair D6.

This is very simple, we just use the variance formula. We already have $\mathbb{E}(X)$, so we can square that for the second term. $\mathbb{E}(X^2)$ is the more annoying one.

$$\begin{aligned}\text{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 &= \sum x^2 P(X = x) - 3.5^2 \\ &= 1^2 \cdot \frac{1}{6} + 2^2 \cdot \frac{1}{6} + 3^2 \cdot \frac{1}{6} + ... + 6^2 \cdot \frac{1}{6} - 3.5^2 = CALC\end{aligned}$$

If we wanted standard deviation, this would just be square rooting the previous value.

2.3 Binomial Distribution

A **Bernouille Trial** is an expreiment where there are only 2 outcomes: Success, and Failure. We say that p is the probability of success.

A **Binomial Experiment** is a series of n bernouille trials.

If we have a random variable X that follows a binomial distribution, then we have the following mean and variance:

$$\mathbb{E}(X) = np \qquad \text{Var}(X) = np(1 - p)$$

The PDF is:

$$P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}$$

Ex. There are 20 samples taken from a process. 1% of samples have problems. X is the number of samples that have problems. What is the probability that the number of samples exceeds its mean by 3 standard deviations?

So we need to get the mean and standard deviations. These can easily be calculated since this follows a binomial distribution:

$$\mathbb{E}(X) = np = 20 \cdot 0.01 = 0.2 \qquad \text{Var}(X) = np(1-p) = 0.2(0.99) = 0.198$$
$$\text{SD}(X) = \sqrt{0.198} = 0.44$$

Now we can find that $P(X > \mathbb{E}(X) + 3\text{SD}(X)) = P(X > 1.535)$. But since we are working with discrete variable, we can only have integers. $P(X > 2) = 1 - P(X \leq 1)$. We use the complement since it is would be very hard to do 18 calculations.

Now we need to actually calculate $1 - P(X \leq 1)$.

$$P(X=0) = \binom{20}{0} 0.01^0 0.99^{20} = CALC$$
$$P(X=1) = \binom{20}{1} 0.01^1 0.99^{19} = CALC$$
$$P(X \leq 1) = 1 - P(X=0) - P(X=1) = CALC$$

2.4 Geometric Distribution

This is a distribution where we want to know the **number of steps before the first success occurs**. Such as the average number of basketball throws needed to score the first point.

We have the following mean and variance:

$$\mathbb{E}(X) = \frac{1}{p} \qquad \text{Var}(X) = \frac{1-p}{p^2}$$

The PDF is:

$$P(X = x) = (1-p)^{x-1}p$$

The negative binomial distribution is similar except we change it to the number of steps before the r**th** success.

$$\mathbb{E}(X) = \frac{r}{p} \qquad \text{Var}(X) = \frac{r(1-p)}{p^2}$$

The PDF is:

$$P(X = x) = \binom{x-1}{r-1}(1-p)^{x-r}p^r$$

2.5 Poisson Distribution

The poisson distribution works off of a rate. We have the parameter λ which is the **number of arrivals** in a **fixed period of time**.

We have the following mean and variance:

$$\mathbb{E}(X) = \lambda \qquad \text{Var}(X) = \lambda$$

The PDF is:

$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

3 Chapter 3: Continuous Random Variables

These are similar to discrete random variables, except for there are **infinite number of outcomes**. They are in a range, so it is not impossible to work with, but it requires different methods.

For example, the heights of a population are not constrained to certain values such as 100lb, 110lb, 120lb. Someone can be 113.223lb. However, there are upper and lower limits such as potentially 5lb up to 500lb. We would not see a human that is 2500lb so we do not even have to consider that part.

If we are given a PDF, we need to integrate over an interval of that PDF in which it basicaly becomes a CDF.

3.1 Expectation

The expectation can be found by integrating the PDF function between the upper and lower limit of the data.

$$\mathbb{E}(h(x)) = \int_{-\infty}^{\infty} h(x)f(x)$$

Ex. We are given the following CDF. Find the mean.

$$F(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ \frac{x}{2} & \text{if } 0 < x < 2 \\ 1 & \text{if } x \geq 2 \end{cases}$$

To get the mean, we need to integrate the PDF. We are given the CDF.

We can get the PDF by taking the derivative of the CDF to get:

$$f(x) = F\prime(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ \frac{1}{2} & \text{if } 0 < x < 2 \\ 1 & \text{if } x \geq 2 \end{cases}$$

Then we can integrate this from 0 to 2:

$$\mathbb{E}(x) = \int_0^2 x \cdot \frac{1}{2} = \frac{1}{2}\left[\frac{x^2}{2}\right]_0^2 = 1$$

Variance and Standard Deviation are defined in the same way as with discrete variables.

3.2 Normal Distribution

This is a continuous distribution that has a specific PDF function.

If a variable Z is normally distributed with **mean of 0** and **variance of 1**, we say $Z \approx \mathcal{N}(0, 1)$.

We represent the normal distribution by Phi (Φ).

$$\Phi(z) = P(Z \leq z)$$

If the mean and variance is not 0 and 1 respectively, we can convert it using the following equation:

$$\frac{x - \mu}{\sigma} = Z \approx N(0, 1)$$

We then have a standard normal table found in the appendix.

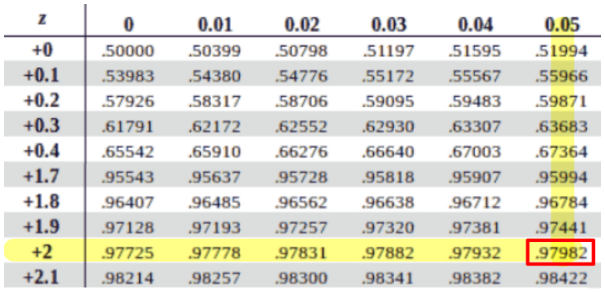
Ex. If we have a mean of 10s to kill a chicken jockey, and a standard deviation of 4s, what is the probability of taking more than 18.2 seconds to kill that pesky chicken jockey?

Notice that we are given the value in terms of standard deviation σ instead of variance σ^2.

We will normalize this value and then we can use the normal value to look it up in the normal table (found in the appendix).

$$\begin{aligned} P(X \geq 18.2) = 1 - P(X \leq 18.2) = 1 - \Phi\left(\frac{18.2 - 10}{4}\right) = 1 - \Phi(2.05) \\ = 1 - 0.9798 = 0.0202 \approx 2\% \end{aligned}$$

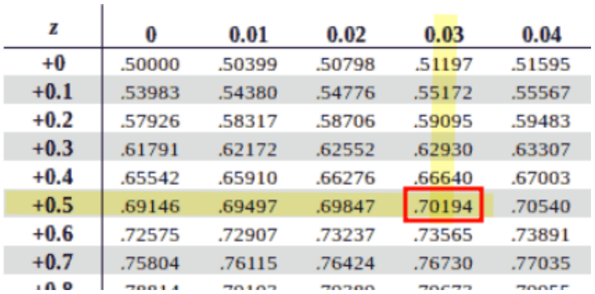
This value was found on the normal table:



z	0	0.01	0.02	0.03	0.04	0.05
+0	.50000	.50399	.50798	.51197	.51595	.51994
+0.1	.53983	.54380	.54776	.55172	.55567	.55966
+0.2	.57926	.58317	.58706	.59095	.59483	.59871
+0.3	.61791	.62172	.62552	.62930	.63307	.63683
+0.4	.65542	.65910	.66276	.66640	.67003	.67364
+1.7	.95543	.95637	.95728	.95818	.95907	.95994
+1.8	.96407	.96485	.96562	.96638	.96712	.96784
+1.9	.97128	.97193	.97257	.97320	.97381	.97441
+2	.97725	.97778	.97831	.97882	.97932	.97982
+2.1	.98214	.98257	.98300	.98341	.98382	.98422

Ex. If we have X normally distributed with mean 10, and variance of 1 ($X \approx \mathcal{N}(10,1)$) then find the value of c such that $P(X \leq c) = 0.7019$.

Here we need to do a bit of reverse engineering. We start with finding out what the normal value (Φ) will be to get 0.7019. I look on the normal table and get 0.53.



z	0	0.01	0.02	0.03	0.04
+0	.50000	.50399	.50798	.51197	.51595
+0.1	.53983	.54380	.54776	.55172	.55567
+0.2	.57926	.58317	.58706	.59095	.59483
+0.3	.61791	.62172	.62552	.62930	.63307
+0.4	.65542	.65910	.66276	.66640	.67003
+0.5	.69146	.69497	.69847	.70194	.70540
+0.6	.72575	.72907	.73237	.73565	.73891
+0.7	.75804	.76115	.76424	.76730	.77035
+0.8	.78814	.79103	.79389	.79673	.79955

Now I know that to get this value of 0.53 I need to apply the normal conversion equation since mean is not 0, and variance is not 1.

$$\frac{c-\mu}{\sigma} = Z \implies \frac{c-10}{\sqrt{1}} = 0.53 \implies c = 10.53$$

Note that in the normal distribution $\mathcal{N}(0,1)$, it uses the **variance** (σ^2) not standard deviation (σ). So that is why we square root the 1 on the denominator.

3.3 Exponential Distribution

This distribution is a poisson proccess with rate of λ.

It is the **waiting time** until the first arrival

We have:

$$f(x) = \lambda e^{-\lambda x} \qquad \mathbb{E}(X) = \frac{1}{\lambda} \qquad \text{Var}(X) = \frac{1}{\lambda^2} \qquad F(x) = 1 - e^{-\lambda x}$$

Ex. In a world, two chicken jockeys spawn per minute. This is a poisson process. What is the probability that we wait more than 1 minute for a chicken jockey to spawn?

We say this is an exponential (or I guess we could do gamma with r=1) distribution since we have poisson variable $\lambda = 2$ and we are finding the waiting time of more than 1 minute, $X > 1$.

$$P(X > 1) = 1 - P(X \leq 1)$$

Then we can use the CDF $F(x)$ of the exponential distribution to solve. We could also use the integral of the PDF (Same thing).

$$= 1 - F(1) = 1 - (1 - e^{-2}) = e^{-2}$$

3.4 Gamma Distribution

This distribution is like the geometric distribution from chapter 2. It is the **waiting time** for the **rth arrival**.

$$f(x) = \frac{x^{r-1}}{(r-1)!}\lambda^r e^{-\lambda x} \qquad \mathbb{E}(X) = \frac{r}{\lambda} \qquad \text{Var}(X) = \frac{r}{\lambda^2}$$

There is no practical way to wright the CDF function for this distribution.

3.5 Joint Distributions

This distribution is where we have 2 variables such as x, and y.

So the PDF function would be $f(x, y)$.

We can work with this similar to the 1 variable functions, except when we have 1 integral/sum, we just need 2 instead. Same idea with expected value, and variance.

4 Chapter 4: Descriptive Statistics and Sampling

To describe a dataset, we have **measures of central tendancy** such as mean and median, and **measures of spread** such as standard deviation, quartiles, and inter-quartile-range.

The median is just the middle value of a **sorted** dataset. If there are 2 middle values, we just take the mean of both of those.

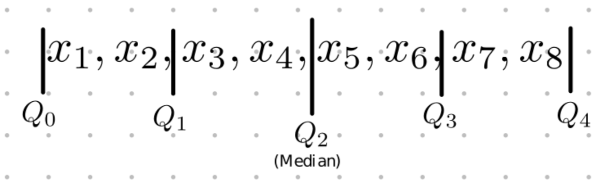
The mean is just:

$$MEAN = \overline{x} = \frac{x_1 + x_2 + ... + x_n}{n}$$

The median is often used since it is not heavily influenced by outliers unlike mean.

4.1 Quartiles

The quartile is like taking the median of the lower half of the data (under the true median).



We call the **Inter Quartile Range (IQR)** as the difference between the third and first quartile $IQR = Q_3 - Q_1$

We identify a datapoint x as an **outlier** if:

$$x < Q_1 - 1.5IQR \qquad \text{or} \qquad x > Q_3 + 1.5IQR$$

4.2 Sample Statistics

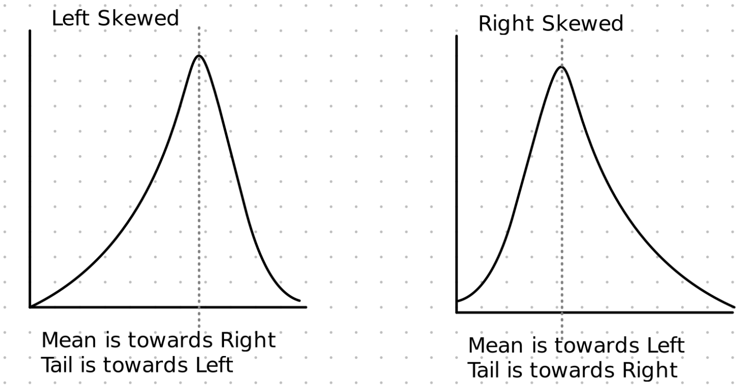
If we do not know the variance of a whole dataset (such as the entire earths population) then we can consider a sample of this population to estimate the population.

We have **sample standard deviation** s and **sample variance** s^2. These estimate the standard deviation σ and variance σ^2 respectively.

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

4.3 Skewness

We call a dataset **left skewed** of the tail of the data is to the left (outliers on the left) or **right skewed** if the tail of the data is on the right (outliers on the right).



4.4 Independant and Identically Distributed (IID) Case

When all variables are independant and identically distributed we say the expected value and variance of the entire set is just the number of variables times the variance/expected value of one item.

$$\mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = n\mu \qquad \text{Var}\left[\sum_{i=1}^{n} X_i\right] = n\sigma^2$$

Then we say that if we are considering a sample of these, we have:

$$\mathbb{E}[\overline{X}] = \mu \qquad \text{Var}[\overline{X}] = \frac{\sigma^2}{n}$$

We can also use the normal distribution if the population is normally distributed to model $\sum_{i=1}^{n} X_i$ or $\overline{X}$.

4.5 Central Limit Theorum

This states that as the **number of runs of an experiment increases**, it will start to reach a **normal distribution**. This is regardless of whether or not the individual experiments are normal or not.

Ex. The average amount of time waiting for a spider jockey to spawn is 8.2 days with standard deviation of 1.5 days. We have a sample of 49 spawn events. What is the probability that we wait less than 10 days for a spider jockey to spawn?

This is very simple. Since it is an IID case where each event is independant, and they all have the same distribution, and n is large, we can use the central limit theorum to approximate it by a normal distribution.

$$P(X \leq 10) = \Phi\left(\frac{10 - 8.2}{1.5/\sqrt{49}}\right) = \Phi(8.4) \approx 1 = 100\%$$

4.6 Difference between 2 Means

We can work with 2 variables $X_1, X_2, ..., X_n$ with μ_1, σ_1^2, and $Y_1, Y_2, ..., Y_m$ with μ_2, σ_2^2 using the following formula:

$$Z = \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}}$$

Ex. We have two random samples. One has size 16, with mean of 75 and standard deviation of 8. The other one has size 9, mean 70, and standard deviation 12. We want to find the probability that their sample means difference is between 3.5 and 5.5.

This is just:

$$3.5 \leq \overline{X}_1 - \overline{X}_2 \leq 5.5$$

I use the normal approximation to add up the variances and means to get:

$$\mu = 75 - 70 = 5 \qquad \sigma^2 = \frac{8^2}{16} + \frac{12^2}{9} = 20$$

And now we can just solve liek a standard normal problem.

$$P(3.5 \le \overline{X}_1 - \overline{X}_2 \le 5.5) = \Phi\left(\frac{5.5-5}{\sqrt{20}}\right) - \Phi\left(\frac{3.5-5}{\sqrt{20}}\right) = 0.1769$$

4.7 Other Distributions

We have 2 other main distributions. The **Chi squared** (χ^2) distribution, and **Student's t** distribution.

Student's t distribution is used when the population variance is unknown, and we have to approximate using the sample variance (standard deviation). This table is found in the appendix.

Both of these distributions have the **degrees of freedom** (v or df) which is usually just $n-1$. This happens since we take away one degree of freedom by estimating the variance.

5 Chapter 5: Point and Interval Estimation

This chapter is mostly about confidence intervals (CI).

We have 2 main confidence intervals that we use. Each of them has an α **value** where if we say the n percent interval, we have $\alpha = 1 - n$.

So for the 95% confidence interval, $\alpha = 0.05$.

The 2 main confidence intervals are the **95 percent, and 99 percent**.

5.1 CI When σ is known

We can use a **normal distribution** to model this since we know σ, n, and $\overline{X}$.

We use the equation:

$$CI = \overline{X} \pm Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Using the normal table we can get $Z_{\alpha/2}$. For $\alpha = 0.05$ we have $Z_{0.025} = 1.96$ and for $\alpha = 0.01$ we have $Z_{0.005} = 2.575$.

Ex. We have a set of data with a standard deviation of 130. We test 9 samples, and we find that there is a sample mean of 4970. We need the 95% confidence interval for this dataset.

We have the following data given from the question:

$$\sigma = 130 \qquad \overline{X} = 4970 \qquad \alpha = 0.05 \qquad n = 9$$

Since σ is known, we can simply sub into the equation with $Z_{\alpha/2} = 1.96$:

$$CI = 4970 \pm 1.96 \cdot \frac{130}{\sqrt{9}} = [4885.07, 5054.93]$$

5.2 CI When σ is unknown

Here we have to find the sample variance s and we know n, and $\overline{X}$.

We use **Student's t distribution** with the equation:

$$CI = \overline{X} \pm t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}}$$

Recall that $n - 1$ is the degrees of freedom for the t distribution.

5.3 CI For a Proportion

When we are dealing with a proportion for a binomial distributions (2 options, either success of failure), we say that P is the probability of success.

We can model this using the **normal distribution** using:

$$CI = P \pm z_{\alpha/2}\sqrt{\frac{P(1-P)}{n}}$$

Ex. In examples, we are not given the variance, or the sample variance. We are however given a probability or the means to calculate a probability.

Then we can just sub the values into the equation and solve.

5.4 Error

We have an error when we estimate using the sample mean $\overline{X}$. This gets smaller as the number of samples n increases. We can find the error using the following equation:

$$E = Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Ex. We have the following values:

$$\sigma = 0.5 \qquad E = 0.2 \qquad \alpha = 0.05 \qquad n = ?$$

We find the sample size required for this error by using the equation:

$$E = Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \implies n = \left(\frac{Z_{\alpha/2}\sigma}{E}\right)^2 = \left(\frac{1.96 \cdot 0.5}{0.2}\right)^2 = 24.01$$

6 Chapter 6: Hypothesis Testing

This chapter is about hypotheses. We create 2 hypotheses. The first one, H_1 is the alternative hypothesis. We test it against the null hypothesis H_0. We do the test for H_0 and we either **reject** the null hypothesis in favour of H_1, or **fail to reject** the null hypothesis. We reject the null if evidence against the null is **strong**.

The Null Hyopothesis H_0 is the claim. It is what we expect to happen. We often want to find evidence to disprove this hypothesis so we can reject it in favour of H_1.

The Alternative Hypothesis H_1 is a different outcome that we are testing for.

6.1 Types of Errors

We commit a **Type 1 error** if we reject H_0 when H_0 is actually true. The probability of a type 1 error is:

$$\alpha = P(\text{reject } H_0 | H_0 \text{ is True})$$

We often use values of $\alpha = 0.01$ or $\alpha = 0.05$.

We commit a **Type 2 error** if we fail to reject H_0 when H_0 is actually false. The probability of a type 2 error is:

$$\beta = P(\text{fail to reject } H_0 | H_0 \text{ is False})$$

6.2 Types of Hypotheses

Typically we are testing if the mean is the same as we expect (null), or if it differs (alternative).

We say that:

$$\begin{aligned} H_0 &: \mu = \mu_0 && \text{Null Hypothesis} \\ H_1 &: \mu \neq \mu_0 && \text{Two Sided Alternative} \\ &\text{OR } \mu < \mu_0 && \text{Left Sided Alternative} \\ &\text{OR } \mu > \mu_0 && \text{Right Sided Alternative} \end{aligned}$$

6.3 Test using P Values

To test using the P value, we need to get the P value, and then compare this to the alpha value.

We use the following procedure:

1. Find H_0, and H_1.
2. Choose the correct α value, usually 0.05 or 0.01
3. For the sample, find z_0 or t_0. Usually found by $z_0 = \frac{\overline{x}-\mu}{\sigma/\sqrt{n}}$ or $t_0 = \frac{\overline{x}-\mu}{S/\sqrt{n}}$ or if it is two sample or proportion, we use the corresponding value.
4. Find the P value by:

$H_1 : \mu > \mu_0$	$P(Z > z_0)$	Or use $P(t(n-1) > t_0)$ for t_0
$H_1 : \mu < \mu_0$	$P(Z < z_0)$	Or use $P(t(n-1) < t_0)$ for t_0
$H_1 : \mu \neq \mu_0$	$2 \cdot \min(P(Z > z_0), P(Z < z_0))$	Or change for t table using t_0

5. Check whether we reject or fail to reject:

   If P value $\leq \alpha$, then reject H_0

   If P value $> \alpha$, then fail to reject H_0

6.4 Test using Confidence Intervals

Rather than find the p value, and then comparing this to the value using α, we can find the confidence interval for the given α and then check if out sample mean is within this range or not.

We take the procedure from the P values, and modify it:

1. Find H_0 and H_1
2. Choose the correct α value, usually 0.05 or 0.01
3. Find the confidence interval for the data: $CI = \overline{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$
4. Check if $\mu \in CI$.

   If $\mu \in CI$, then we fail to reject H_0

   If $\mu \notin CI$, then we reject H_0

Ex. We have the following data.

$$n = 9 \qquad \sigma^2 = 25 \qquad \overline{X} = 23 \qquad \alpha = 0.05$$
$$H_0 : \mu = 20 \qquad H_1 : \mu \neq 20$$

So we want to see whether or not to reject the claim that the average is 20.

We will create a confidence interval and see if the mean of 20 is within this confidence interval (fail to reject) or falls outside this CI (reject).

$$CI = 23 \pm 1.96 \cdot \frac{5}{3} = [19.73, 26.26]$$

Since $\mu \in CI$, then we fail to reject the claim that $\mu = 20$.

6.5 Test for a Proportion

We can test using a proportion in a very similar way to the standard tests using variance, except we use the equation:

$$z_0 = \frac{x - np}{\sqrt{np(1-p)}}$$

6.6 Two Sample Test

If we have two samples; one of them is before an event and one after an event, and we want to test a hypothesis about this difference D_i, then we do:

$$t_0 = \frac{\overline{D}}{S_D/\sqrt{n}}$$

We can get the sample variance S_D^2 and average difference $\overline{D}$ through:

$$\overline{D} = \frac{1}{n}\sum D_i \qquad S_D^2 = \frac{1}{n-1}\sum (D_i - \overline{D})^2$$

Ex. We have 10 people who are in a program to reduce their weight. Before and after statistics as well as difference is shown below. Using 95% confidence, is this program effective?

Before	195	213	247	201	187	210	215	246	294	310
After	187	195	221	190	175	197	199	221	278	285
Difference	-8	-18	-26	-11	-12	-13	-16	-25	-16	-25

Now we need to follow the standard steps to find the p-value.

The hypotheses are: $H_0 : \mu_0 = 0 \qquad H_1 : \mu_0 < 0$

We have $\alpha = 0.05$

To get t_0, we will use the two sample test equation. We need to find the sample variance S^2 and the sample mean $\overline{D}$.

I will use a spreadsheet to calculate these using the equations:

$$\overline{D} = \frac{1}{n}\sum D_i = -17 \qquad S_D^2 = \frac{1}{n-1}\sum (D_i - \overline{D})^2 = 41.11$$

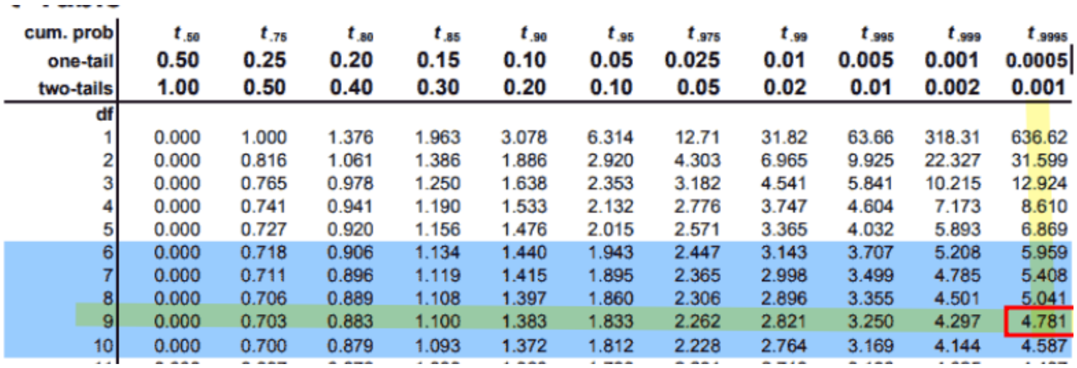
And now I can actually calculate t_0.

$$t_0 = \frac{\overline{D}}{S_D/\sqrt{n}} = \frac{-17}{\sqrt{41.11}/\sqrt{10}} = -8.38$$

Now we find the P value using the second row of the table in step 4.

$$P(t(10-1) < -8.38) < 0.0005$$

We could not get the exact value on the t table, but we see that it is larger than even the highest value (4.781) so it gives us a value smaller than 0.0005.



cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587

The final step asks us to compare the P value (less than 0.0005) to the alpha value, and it is smaller. So re reject H_0.

This means we reject the claim that the mean difference is 0 (no effect of program) in favour of the claim that the mean difference is less than 0 (good effect of the weight loss program).

7 Chapter 7: Linear Regresion

7.1 Correlation Coefficient

The **coefficient of correlation** ρ between 2 variables x and y is:

$$\rho_{xy} = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum(x_i - \overline{x})^2 \sum(y_i - \overline{y})^2}} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$

We can also wright these as:

$$S_{xy} = \sum x_i y_i - n\overline{x}\overline{y}$$
$$S_{xx} = \sum x_i^2 - n\overline{x}^2$$
$$S_{yy} = \sum y_i^2 - n\overline{y}^2$$

7.2 Linear Regression

We can get a line of best fit using the equation of:

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

Here ϵ is the error term which is often omitted.

The β_1 can be found using the following equation:

$$\beta_1 = \frac{S_{xy}}{S_{xx}}$$

The β_0 can be found by subbing in a known values for the equation of $\overline{y} = \beta_0 + \beta_1\overline{x}$ and solving for β_0.

We can also estimate the **variance** by doing:

$$\hat{\sigma}^2 = \frac{S_{yy} - \beta_1 S_{xy}}{n-2}$$

Ex. For these examples we are typically given a lot of values such as:

$$\sum x_i = ... \qquad \sum y_i = ... \qquad \sum x_i^2 = ... \qquad \sum x_i y_i = ... \qquad \sum y_i^2 = ... \qquad n = ...$$

Then we need to find the equation for linear regression.

We know how to get $\beta_1 = \frac{S_{xy}}{S_{xx}}$.

Finding these values such as S_{xx} we get since we know the $\sum x_i^2$, and while we do not directly know $\overline{x}$, we can find it by definition through $\overline{x} = \frac{1}{n}\sum x_i$

$$S_{xx} = \sum x_i^2 - n\overline{x}^2 = \sum x_i^2 - n\left(\frac{1}{n}\sum x_i\right)^2$$

The same idea applies for S_{xy}

$$S_{xy} = \sum x_i y_i - n\frac{\sum x_i}{n}\frac{\sum y_i}{n}$$

And then we have β_1.

To find β_0, we can sub into the equation of $y = \beta_0 + \beta_1 x \implies \overline{y} = \beta_0 + \beta_1\overline{x}$ since we know $\overline{y}$ and $\overline{x}$.

7.3 Hypothesis Testing

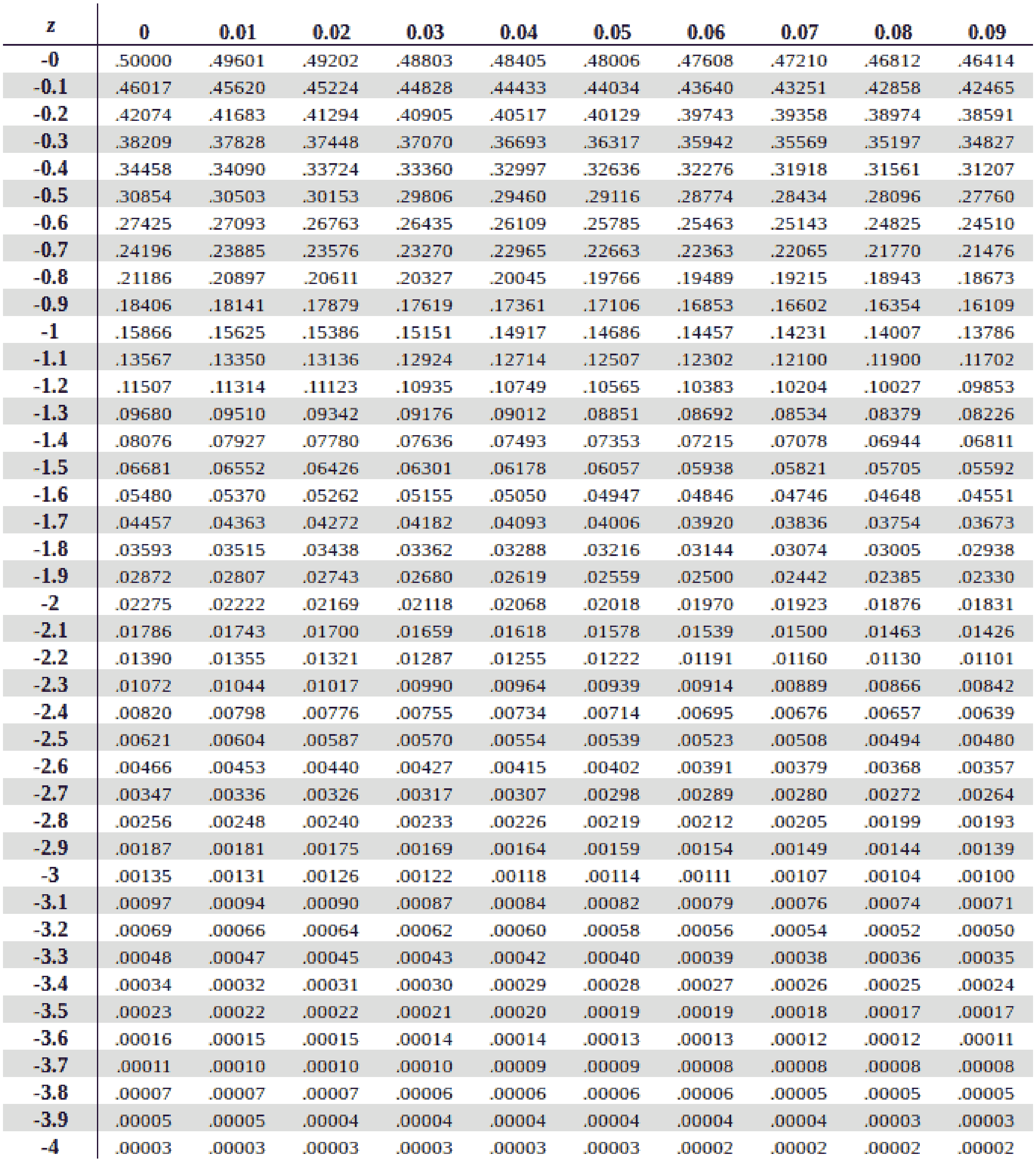
We can do hypothesis testing using these β values such as where:

$$H_0 : \beta_0 = \beta_{0,0} \qquad H_1 : \beta_0 \neq \beta_{0,0}$$

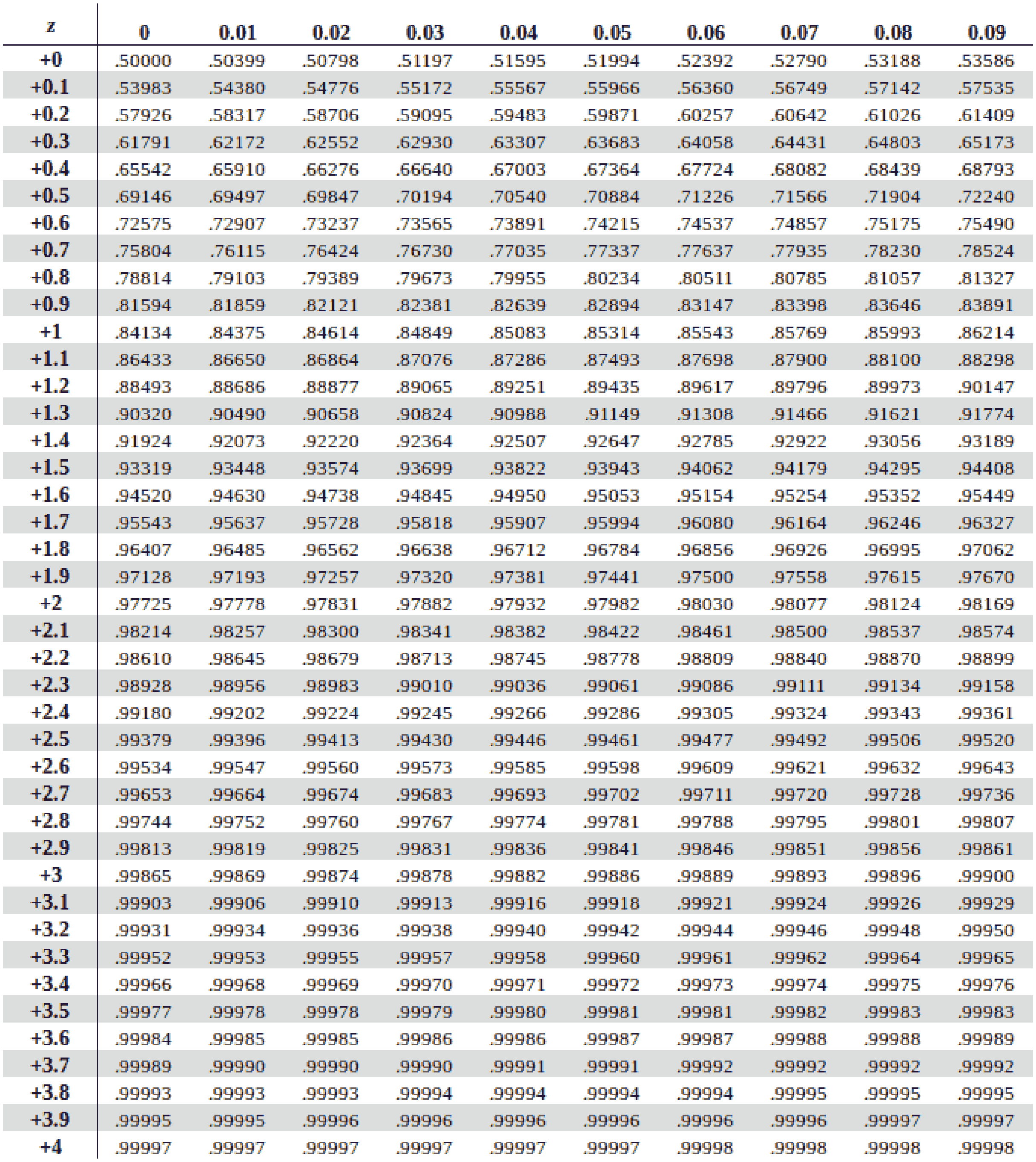
Similarly to chapter 6, this result is often **normally distributed** so can be easily calculated using the normal table or Student's t table.

8 Appendix

8.1 Normal Distribution Table (Z-Table)



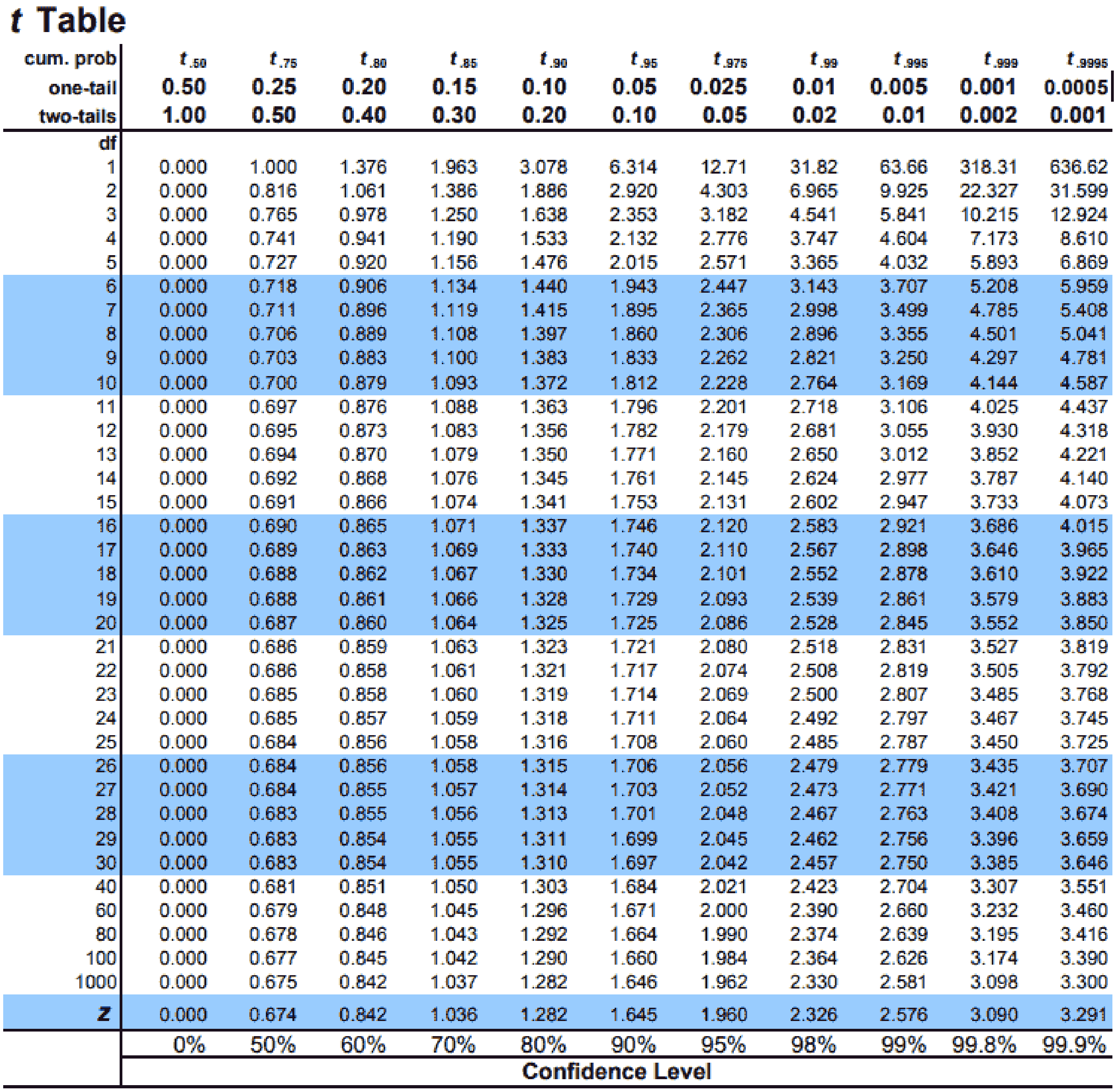
z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-0	.50000	.49601	.49202	.48803	.48405	.48006	.47608	.47210	.46812	.46414
-0.1	.46017	.45620	.45224	.44828	.44433	.44034	.43640	.43251	.42858	.42465
-0.2	.42074	.41683	.41294	.40905	.40517	.40129	.39743	.39358	.38974	.38591
-0.3	.38209	.37828	.37448	.37070	.36693	.36317	.35942	.35569	.35197	.34827
-0.4	.34458	.34090	.33724	.33360	.32997	.32636	.32276	.31918	.31561	.31207
-0.5	.30854	.30503	.30153	.29806	.29460	.29116	.28774	.28434	.28096	.27760
-0.6	.27425	.27093	.26763	.26435	.26109	.25785	.25463	.25143	.24825	.24510
-0.7	.24196	.23885	.23576	.23270	.22965	.22663	.22363	.22065	.21770	.21476
-0.8	.21186	.20897	.20611	.20327	.20045	.19766	.19489	.19215	.18943	.18673
-0.9	.18406	.18141	.17879	.17619	.17361	.17106	.16853	.16602	.16354	.16109
-1	.15866	.15625	.15386	.15151	.14917	.14686	.14457	.14231	.14007	.13786
-1.1	.13567	.13350	.13136	.12924	.12714	.12507	.12302	.12100	.11900	.11702
-1.2	.11507	.11314	.11123	.10935	.10749	.10565	.10383	.10204	.10027	.09853
-1.3	.09680	.09510	.09342	.09176	.09012	.08851	.08692	.08534	.08379	.08226
-1.4	.08076	.07927	.07780	.07636	.07493	.07353	.07215	.07078	.06944	.06811
-1.5	.06681	.06552	.06426	.06301	.06178	.06057	.05938	.05821	.05705	.05592
-1.6	.05480	.05370	.05262	.05155	.05050	.04947	.04846	.04746	.04648	.04551
-1.7	.04457	.04363	.04272	.04182	.04093	.04006	.03920	.03836	.03754	.03673
-1.8	.03593	.03515	.03438	.03362	.03288	.03216	.03144	.03074	.03005	.02938
-1.9	.02872	.02807	.02743	.02680	.02619	.02559	.02500	.02442	.02385	.02330
-2	.02275	.02222	.02169	.02118	.02068	.02018	.01970	.01923	.01876	.01831
-2.1	.01786	.01743	.01700	.01659	.01618	.01578	.01539	.01500	.01463	.01426
-2.2	.01390	.01355	.01321	.01287	.01255	.01222	.01191	.01160	.01130	.01101
-2.3	.01072	.01044	.01017	.00990	.00964	.00939	.00914	.00889	.00866	.00842
-2.4	.00820	.00798	.00776	.00755	.00734	.00714	.00695	.00676	.00657	.00639
-2.5	.00621	.00604	.00587	.00570	.00554	.00539	.00523	.00508	.00494	.00480
-2.6	.00466	.00453	.00440	.00427	.00415	.00402	.00391	.00379	.00368	.00357
-2.7	.00347	.00336	.00326	.00317	.00307	.00298	.00289	.00280	.00272	.00264
-2.8	.00256	.00248	.00240	.00233	.00226	.00219	.00212	.00205	.00199	.00193
-2.9	.00187	.00181	.00175	.00169	.00164	.00159	.00154	.00149	.00144	.00139
-3	.00135	.00131	.00126	.00122	.00118	.00114	.00111	.00107	.00104	.00100
-3.1	.00097	.00094	.00090	.00087	.00084	.00082	.00079	.00076	.00074	.00071
-3.2	.00069	.00066	.00064	.00062	.00060	.00058	.00056	.00054	.00052	.00050
-3.3	.00048	.00047	.00045	.00043	.00042	.00040	.00039	.00038	.00036	.00035
-3.4	.00034	.00032	.00031	.00030	.00029	.00028	.00027	.00026	.00025	.00024
-3.5	.00023	.00022	.00022	.00021	.00020	.00019	.00019	.00018	.00017	.00017
-3.6	.00016	.00015	.00015	.00014	.00014	.00013	.00013	.00012	.00012	.00011
-3.7	.00011	.00010	.00010	.00010	.00009	.00009	.00008	.00008	.00008	.00008
-3.8	.00007	.00007	.00007	.00006	.00006	.00006	.00006	.00005	.00005	.00005
-3.9	.00005	.00005	.00004	.00004	.00004	.00004	.00004	.00004	.00003	.00003
-4	.00003	.00003	.00003	.00003	.00003	.00003	.00002	.00002	.00002	.00002



z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
+0	.50000	.50399	.50798	.51197	.51595	.51994	.52392	.52790	.53188	.53586
+0.1	.53983	.54380	.54776	.55172	.55567	.55966	.56360	.56749	.57142	.57535
+0.2	.57926	.58317	.58706	.59095	.59483	.59871	.60257	.60642	.61026	.61409
+0.3	.61791	.62172	.62552	.62930	.63307	.63683	.64058	.64431	.64803	.65173
+0.4	.65542	.65910	.66276	.66640	.67003	.67364	.67724	.68082	.68439	.68793
+0.5	.69146	.69497	.69847	.70194	.70540	.70884	.71226	.71566	.71904	.72240
+0.6	.72575	.72907	.73237	.73565	.73891	.74215	.74537	.74857	.75175	.75490
+0.7	.75804	.76115	.76424	.76730	.77035	.77337	.77637	.77935	.78230	.78524
+0.8	.78814	.79103	.79389	.79673	.79955	.80234	.80511	.80785	.81057	.81327
+0.9	.81594	.81859	.82121	.82381	.82639	.82894	.83147	.83398	.83646	.83891
+1	.84134	.84375	.84614	.84849	.85083	.85314	.85543	.85769	.85993	.86214
+1.1	.86433	.86650	.86864	.87076	.87286	.87493	.87698	.87900	.88100	.88298
+1.2	.88493	.88686	.88877	.89065	.89251	.89435	.89617	.89796	.89973	.90147
+1.3	.90320	.90490	.90658	.90824	.90988	.91149	.91308	.91466	.91621	.91774
+1.4	.91924	.92073	.92220	.92364	.92507	.92647	.92785	.92922	.93056	.93189
+1.5	.93319	.93448	.93574	.93699	.93822	.93943	.94062	.94179	.94295	.94408
+1.6	.94520	.94630	.94738	.94845	.94950	.95053	.95154	.95254	.95352	.95449
+1.7	.95543	.95637	.95728	.95818	.95907	.95994	.96080	.96164	.96246	.96327
+1.8	.96407	.96485	.96562	.96638	.96712	.96784	.96856	.96926	.96995	.97062
+1.9	.97128	.97193	.97257	.97320	.97381	.97441	.97500	.97558	.97615	.97670
+2	.97725	.97778	.97831	.97882	.97932	.97982	.98030	.98077	.98124	.98169
+2.1	.98214	.98257	.98300	.98341	.98382	.98422	.98461	.98500	.98537	.98574
+2.2	.98610	.98645	.98679	.98713	.98745	.98778	.98809	.98840	.98870	.98899
+2.3	.98928	.98956	.98983	.99010	.99036	.99061	.99086	.99111	.99134	.99158
+2.4	.99180	.99202	.99224	.99245	.99266	.99286	.99305	.99324	.99343	.99361
+2.5	.99379	.99396	.99413	.99430	.99446	.99461	.99477	.99492	.99506	.99520
+2.6	.99534	.99547	.99560	.99573	.99585	.99598	.99609	.99621	.99632	.99643
+2.7	.99653	.99664	.99674	.99683	.99693	.99702	.99711	.99720	.99728	.99736
+2.8	.99744	.99752	.99760	.99767	.99774	.99781	.99788	.99795	.99801	.99807
+2.9	.99813	.99819	.99825	.99831	.99836	.99841	.99846	.99851	.99856	.99861
+3	.99865	.99869	.99874	.99878	.99882	.99886	.99889	.99893	.99896	.99900
+3.1	.99903	.99906	.99910	.99913	.99916	.99918	.99921	.99924	.99926	.99929
+3.2	.99931	.99934	.99936	.99938	.99940	.99942	.99944	.99946	.99948	.99950
+3.3	.99952	.99953	.99955	.99957	.99958	.99960	.99961	.99962	.99964	.99965
+3.4	.99966	.99968	.99969	.99970	.99971	.99972	.99973	.99974	.99975	.99976
+3.5	.99977	.99978	.99978	.99979	.99980	.99981	.99981	.99982	.99983	.99983
+3.6	.99984	.99985	.99985	.99986	.99986	.99987	.99987	.99988	.99988	.99989
+3.7	.99989	.99990	.99990	.99990	.99991	.99991	.99992	.99992	.99992	.99992
+3.8	.99993	.99993	.99993	.99994	.99994	.99994	.99994	.99995	.99995	.99995
+3.9	.99995	.99995	.99996	.99996	.99996	.99996	.99996	.99996	.99997	.99997
+4	.99997	.99997	.99997	.99997	.99997	.99997	.99998	.99998	.99998	.99998

Tables taken from https://www.ztable.net/

8.2 Student's t Table



t Table

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	**0.50**	**0.25**	**0.20**	**0.15**	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**	**0.001**	**0.0005**
two-tails	**1.00**	**0.50**	**0.40**	**0.30**	**0.20**	**0.10**	**0.05**	**0.02**	**0.01**	**0.002**	**0.001**
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence Level										

Table taken from https://www.tdistributiontable.com/