Contents

1 Estimators and confidence intervals

1.1 Estimators

Lets say we have some data and there is an unknown parameter like the mean or variance of where the data came from. For example, lets say the data is

$$\{8, 8, 9, 13, 15\}$$

In this case, the sample mean is the mean of the data, which is 10.6, and if the data is from a random variable X we would call this $\bar{X}$. However, the actual mean, which is the unknown quantity, is what we call μ. We say that 10.6, which is the sample mean, is an estimator for μ.

Definition. An estimator for a parameter is any function of the data. However, most functions are silly estimators.

Definition. An estimator for a parameter θ, denoted $\hat{\theta}$, is said to be unbiased if we know that $E\left(\hat{\theta}\right) = \theta$.

Example. In the above example, taking the first data point we get, X_1, as our estimator would technically give an unbiased estimator because $E(X_1) = \mu$, however we will get a much more accurate estimate if we take our estimator to me $\bar{X}$, which is still unbiased and has a lower error in general.

The other thing to know about estimators is that given some data, one can estimate the variance. You have to be a bit careful, because you cannot just take the mean of $\left(X_i - \overline{X}\right)^2$. The reason is that in general, $\left(X_i - \overline{X}\right)^2 \neq (X_i - \mu)^2$. It turns out that you take an unbiased estimator for the variance, when n is the number of data points, if instead of calculating $\frac{1}{n}\sum\left(X_i - \overline{X}\right)^2$ (which generally underestimates the variance), you calculate $\frac{1}{n-1}\sum\left(X_i - \overline{X}\right)^2$. In level 6.3 you will see a really nice visual argument that shows why this is unbiased and not just naively making it slightly larger to make up for the fact that it is an underestimate.

1.2 Confidence intervals

Definition. A confidence interval for a parameter θ is an interval of numbers that we generate based on data, and we say it is an $X\%$ confidence interval if on average if we keep generating them, $X\%$ of the intervals will contain θ. What this is not saying is that given a confidence interval, θ has an $X\%$ chance to lie in it. The difference is subtle so I will show an example.

Example. Given a uniform distribution from $\theta - \frac{1}{2}$ to $\theta + \frac{1}{2}$, if we generate 2 numbers from this, then θ lies between them 50% of the time so if we take our confidence interval to be from the min to the max of the two observations, we get a 50% confidence interval. However if we roll 0.2 and 0.9 we know θ **must** lie between them, so that is the subtle difference.

Example. We will use the fact that given a normal distribution 95% of the data will fall within about 1.96 standard deviations of the mean. Let X be normally distributed with known variance σ and unknown mean μ, then if we generate 1 data point X, a 95% confidence interval for μ would be $X \pm 1.96\sigma$ because about 95% of these intervals will contain μ.

2 More distributions

2.1 Geometric

Imagine we have an event which occurs with probability p and we want to consider how many trials it takes for the event to occur. Assume trials are independent and all that.

The probability of it occuring on trial 1 is of course p. The probability of it occuring on trial n is equal to the probability it does not occur in the first n-1 trials times the probability that it occurs on the n'th trial, which is $p\left(1-p\right)^{n-1}$.

Proposition. As is expected from the definition, the mean of a geometric distribution is $\frac{1}{p}$.

Proof. We just need to find

$$\sum_{n=1}^{\infty}\left(nP(x=n)\right)=\sum_{n=1}^{\infty}np\left(1-p\right)^{n-1}=p\sum_{n=1}^{\infty}n\left(1-p\right)^{n-1}=p\frac{d}{d\left(1-p\right)}\sum_{n=1}^{\infty}\left(1-p\right)^{n}$$

In level 4 we proved we can differentiate power series inside the radius of convergence which in this case is 1, ie we are fine.

$$=-p\frac{d}{dp}\sum_{n=1}^{\infty}\left(1-p\right)^{n}=-p\frac{d}{dp}\frac{1-p}{1-\left(1-p\right)}=-p\frac{d}{dp}\frac{1-p}{p}=-p*\left(\frac{-1}{p^{2}}\right)=\frac{1}{p}$$

by the geometric series formula.

□

2.2 Negative binomial

The negative binomial with parameters r and p is like the geometric distribution but instead of measuring how long it will take to get an event, we are measuring how long it will take to get r events where the probability is p.

To find $P\left(X=n\right)$ we note that this happens exactly under 2 conditions:

- In the first n-1 trials we get exactly r-1 successes, we can calculate the probability of this using the binomial formula

- On the n'th trial we get a success, this has probability p.

We therefore can multiply these together. We get

$$P\left(X=n\right)=\left(\begin{array}{c}n-1\\r-1\end{array}\right)p^{r}\left(1-p\right)^{n-r}$$

Given a question like "Show that it takes more than X trials" just calculate the probability that there are **at most** r-1 successes in the first X trials. Given a question like "Find the probability we get the r'th success on the n'th trial given

that the first one was successful", treat it like a problem where the first one is removed and we need to find $P(X = n-1)$ where the parameters for the negative binomial are p and r-1.

Note that since a negative binomial is the sum of r geometric distributions, its mean is $\frac{r}{p}$.

It is annoying to prove the variance of a geometric distribution, but given this and the fact that variances add we will get that the variance of a geometric distribution is $\frac{1-p}{p^2}$ and the variance of a negative binomial is $\frac{r(1-p)}{p^2}$. We will show all this in level 6.3.

3 Probability generating functions

A probability generating function (pgf) is defined as follows: Suppose X is a random variable that follows a distribution that takes only non-negative integers, then we define the probability generating function of X as a function of t as follows:

$$\sum_{n=0}^{\infty} P(X = n) t^n$$

Example. if X follows a distribution such that $P(X = 1) = 0.1$, $P(X = 2) = 0.2$, $P(X = 3) = 0.3$, $P(X = 4) = 0.4$, then its pgf is $0.1t + 0.2t^2 + 0.3t^3 + 0.4t^4$.

Ok enough of that lets find the pgf of some standard distributions.

The pgf of a binomial distribution with parameters (n, p) is as follows (by carefully applying the binomial theorem):

$$\sum_{r=0}^{n} \left(\begin{array}{c} n \\ r \end{array} \right) p^r (1-p)^{n-r} t^r = (1-p)^n \sum_{r=0}^{n} \left(\begin{array}{c} n \\ r \end{array} \right) \left(\frac{pt}{1-p} \right)^r = (1-p)^n \left(1 + \frac{pt}{1-p} \right)^n = (1 - p + pt)^n$$

The pgf of a poisson distribution with parameter λ is as follows (by applying Taylor series):

$$\sum_{n=0}^{\infty} \frac{e^{-\lambda} \lambda^n}{n!} t^n = e^{-\lambda} \sum_{n=0}^{\infty} \frac{(\lambda t)^n}{n!} = e^{-\lambda} e^{\lambda t} = e^{(t-1)\lambda}$$

The pgf of a geometric distribution with parameter p is as follows (by applying Geometric series, valid if $|t| \leq 1$ since $p < 1$):

$$\sum_{n=1}^{\infty} p (1-p)^{n-1} t^n = \frac{p}{(1-p)} \sum_{n=1}^{\infty} ((1-p) t)^n = \frac{p}{(1-p)} \frac{(1-p) t}{1 - (1-p) t} = \frac{pt}{1 - (1-p) t}$$

We will see the pgf of a negative binomial shortly.

Note that if the radius of convergence of the pgf is > 1 then its derivative (by level 4) is

$$\sum_{n=0}^{\infty} nP(X = n) t^n$$

Evaluated when $t = 1$ this gives $\sum_{n=0}^{\infty} nP(X = n) = E(X)$. Therefore we can find the expectation by differentiating the pgf and evaluating it at 1.

It will follow from level 6.1 that the radius of convergence must be at least 1: We will show that power series essentially converge in a circle in the complex plane, and since the pgf is $\sum_{n=0}^{\infty} P(X = n)$ which converges 1 when $t = 1$, then the result will follow. In level 6.3 we will show that if g is the pgf, we will always get the expectation by taking $\lim_{h \to 0} g'(1 - h)$.

Note that pgf's and all their derivatives are increasing as t increases since each term in the sum clearly is since the probabilities are positive. This means pgfs always curve upwards, and they are 1 when $t = 1$.

With those technical details aside, note that the second derivative of a pgf is $\sum_{n=0}^{\infty} n(n-1) P(X = n) t^n = E[X(X-1)]$.

It follows that

$$g''(1) + g'(1) - (g'(1))^2 = E[X(X-1)] + E[X] - E[X]^2 = E[X^2] - E[X]^2 = Var(X)$$

Actually, we can use this formula to get all the variance results that w will prove more directly in level 6.3. For example, the binomial has $g = (1-p+pt)^n$ so $g' = np(1-p+pt)^{n-1}$ and $g'' = n(n-1)p^2(1-p+pt)^{n-2}$. g is just a polynomial so its radius of convergence is infinite. When $t = 1$, we have $g'(1) = np$ (Excatly the mean!) and $g''(1) = n(n-1)p^2$. Now

$$g''(1) + g'(1) - (g'(1))^2 = n(n-1)p^2 - np - (np)^2 = np(1-p)$$

where at the end I expanded and simplified. So yeah that is a neat way to derive it, but like I said we will do a direct derivation in level 6.3.

Now, one last property: Given a random variable X that takes non-negative integers with probability generating function $g(t)$, if A and B are positive integers then the probability generating function of $Ax + B$ is $t^B g\left(t^A\right)$. It is not difficult to see why this is the case, but I will put the explanation in level 6.3.

Also, in level 6.3 we will show (It's a really neat argument! And for technical details about infinite sums that we use in the argument these will be justified in level 6.1) that if $X_1, X_2, ..., X_n$ are independent random variables with pgf $g(t)$, then the pgf of $X_1 + X_2 + ... + X_n$ is $g(t)^n$.

From this result, it follows that the pgf of a negative binomial with parameters p and r, from the known pgf of the geometric distribution, is $\left(\frac{pt}{1-(1-p)t}\right)^r$.

4 Properties of hypothesis tests

4.1 Types of errors

Recall that H_0 is the "null" hypothesis, ie the "neutral" hypothesis. Suppose you carry out a hypothesis test. Then there are two ways it can go wrong. Before reading ahead, try to guess what they are (Hint: It is the simplest most obvious answer)!

Definition. A type I error is when you reject H_0 but H_0 is true and you just got "lucky" (or unlucky, depending on the situation).

Definition. A type II error is when you accept H_0 but it is false. For example, this would happen if Dream was cheating in Minecraft but still got luck within the normal range.

Definition. The size of a hypothesis test is the probability of a type I error. This is just the significance level. We have seen in level 3.3 that this is the "actual" significance level, which for discrete data will be different from the significance level that we declare we will work with at the start.

4.2 Power of a test

Definition. The power of a hypothesis test is the probability, given that H_0 is untrue and given the corrected parameter, that the test will be correct (ie, not give a type II error).

Example. Suppose we want to test whether $X \sim N(30, 5^2)$ and X represents some real world situation. At the 5% level of significance, our critical values would be 20.2 and 39.8. If we are given that H_0 is false and the mean is actually 40, then $X \sim N\left(40, 5^2\right)$ and the power of the test is $1 - P(20.2 < X < 39.8)$. Based on a statistical table, this is about 0.5160, so in this case not very powerful as the actual significance level was only 2 standard deviations out. This agrees with the intuitive idea that it doesn't really make sense to test X by just rolling it once and we want to roll it more than once - We will see more on this idea shortly, in fact in the next section.

5 Distribution of sample mean and law of large numbers

Let X be a random variable with mean μ and variance σ^2. Then if $X_1, X_2, ..., X_n$ are independent copies of X, then by additivity of mean and variance, $E[X_1 + X_2 + ... + X_n] = n\mu$ and $Var[X_1 + X_2 + ... + X_n] = n\sigma^2$. But recall that variance is square-linear in the sence that $Var[cX] = c^2 Var[X]$. Therefore, $Var\left[\frac{X_1+X_2+...+X_n}{n}\right] = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$.

This formalizes the idea that if you take many random variables the variance of the mean will get smaller, ie you will have

a better idea of the mean. This is one variant of the **law of large numbers**. See the probability tripos course (level 8.4) for more details.

6 Properties of the normal distribution

6.1 Sums of normal random variables

It turns out - by a nice geometric argument we will do in level 6.3 (the same argument that 3Blue1Brown did in his video on this topic) - that if $X \sim N\left(\mu_1, \sigma_1^2\right)$ and $Y \sim N\left(\mu_2, \sigma_2^2\right)$ and X and Y are independent then $X+Y \sim N\left(\mu_1+\mu_2, \sigma_1^2+\sigma_2^2\right)$. We know (since mean and variance are additive) that this has the correct mean and variance, and the surprising part is that it is actually a normal.

Example. Let $X \sim N\left(1, 0.2^2\right)$ and suppose we have 100 identical independent copies of X and we take the mean $\bar{x}$, then what is $P\left(\bar{x} > 1.04\right)$?

The solution is by first observing that since X has a normal distribution, that so does $\bar{x}$ since it is a sum of multiples of instances of x. We also know that the variance of X is 0.2^2 so the variance of $\bar{x}$ is $\frac{0.2^2}{100}$. The square root of this is 0.02 so that is the standard deviation of $\bar{x}$. Therefore, 1.04 is 2 standard deviations out. In a normal distribution, by a table, the probability we are at least 2 standard deviations out is about 2.28% so that is our answer.

Note that if I give n observations from a normal distribution with standard deviation σthen my 95% confidence interval for the mean, which before was $X \pm 1.96\sigma$ now becomes more narrow, specifically the mean is effectively a normal distirbution with a variance $\frac{\sigma^2}{n}$ so a 95% confidence interval for the mean becomes $\bar{X} \pm 1.96\frac{\sigma}{\sqrt{n}}$.

6.2 The central limit theorem

The central limit theorem is very difficult to prove, but we essentially saw it in level 3.3 when we were talking about approximating other distributions by noraml distributions. However, we will do it in level 6.3 (using level 6.2 as background). Anyway, it says that if we let $X_1, X_2, ..., X_n$ be copies of *any* distribution with variance σ^2 and mean μ(Both have to exist, so it is not about every distribution!), then if we add them together, shift it so that the mean is 0, and scale it so that the variance is 1, ie we take

$$\frac{X_1 + X_2 + ... + X_n}{\sigma\sqrt{n}} - \mu$$

and call this Y, then as n gets large this will always approach an $N\left(0,1\right)$ distribution, in the sense that $P\left(Y \leq k\right)$ as a sequence depending on n will approach the value that it would be if Y were $N\left(0,1\right)$.

One reason you might expect this to be true is that whatever distribution Y converges to, if anything must have the property that $Y+Y$ has a similar distribution just scaled differently, and we know this is true of the normal.

Based on numerical data (ie, this is not a theorem which we can prove just something that tends to be the case), around $n = 30$ gives a good approximation, less or more depending on how bell shaped the distribution is to begin with, and convergence is faster near the peak of the distribution and slower near the tails.

Example. Since a $Po\left(100\right)$ is the sum of 100 $Po\left(1\right)$ distributions, it is well approximated by an $N\left(100, 100\right)$.

Example. Since a $B\left(n,p\right)$ is the sum of n $B\left(1,p\right)$ distributions (Which, by the way, are called Bernoulli distributions with parameter p), it is well approximated by a normal distribution with the right mean and variance. In general, the approximation works if np and $n\left(1-p\right)$ are both not too small (so that the distribution has room on both sides of the mean), and you will often see $p \approx 0.5$ as the condition.

Even if the starting distribution is unknown, you can still use the theorem to solve problems (provided you know that the variance exists).

For those interested,

Sketch of Level 6.3 proof. We will define the characteristic function of a distribution of X as $\phi_X\left(t\right) := E\left(e^{itX}\right)$. We will show (and this will be very difficult - ChatGPT once told me you can't do it with A level techniques which I proved wrong on the technicality that I never said it had to be short) that this determines the distribution, that it is sufficiently

well behaved, and (the easier part) that the characteristic function of the standardized sample mean converges to the characteristic function of a normal distribution (which we will compute).

7 Chi squared tests

Hehe this is the thing I was stuck on proving why it works for so long and then ended up coming up with a new proof of it (which as far as I know has never been discovered before) which I present in level 6.3. It is very complicated. I also have an even more complicated and clunky proof of it in the misc results section, that assumes the central limit theorem. For now we will assume that the procedure works.

Ok so here is the procedure. Lets say that we want to test if X follows a certain distribution. In this example, we roll a dice 60 times and we want to test if it is fair. Suppose that the table of frequencies is as follows:

1	2	3	4	5	6
11	8	10	13	9	9

These are the "Observed counts". The "Expected counts" under the null hypothesis would be $(10, 10, 10, 10, 10, 10)$.

In this case, there are 6 cells, and we can calculate

$$\sum_{i=1}^{6} \frac{(O_i - E_i)^2}{E_i} = \frac{(11-10)^2}{10} + \frac{(8-10)^2}{10} + \frac{(10-10)^2}{10} + \frac{(13-10)^2}{10} + \frac{(9-10)^2}{10} + \frac{(9-10)^2}{10} = 1.6$$

1.6 is the "Chi squared statistic".

Now I will define (I will not actually define this properly) the number of "degrees of freedom". In the above example, there are 5 degrees of freedom since there are 6 cells and 1 constraint, the constraint being that the total of all the cell values is 60.

Now, here is the magic part (which we will prove in level 6.3 - and I warn you the proof is not for the faint of heart, as I said at the beginning of this section the proof is a new discovery to my knowledge as previous proofs were even more complicated and require a math degree, I keep going on about this because of how long I spent obsessing over this because I refuse to use things that feel like magic haha). If each cell has a null hypothesis probability (in this case $\frac{1}{6}$ for all of them), then as n (the total number of trials, 60 in the example above) gets large, the statistic $\sum_{i=1}^{6} \frac{(O_i-E_i)^2}{E_i}$ approaches a distribution from a family of distributions called the chi squared distributions, and which one it approaches depends only on the number of degrees of freedom.

In the above case, we are looking for the "5" member in the family of chi squared distributions since there are 5 degrees of freedom.

Figure 1 shows the percentage points of the chi squared distribution, the number of degrees of freedom in the leftmost column.

	P										
DF	0.995	0.975	0.20	0.10	0.05	0.025	0.02	0.01	0.005	0.002	0.001
1	0.0000393	0.000982	1.642	2.706	3.841	5.024	5.412	6.635	7.879	9.550	10.828
2	0.0100	0.0506	3.219	4.605	5.991	7.378	7.824	9.210	10.597	12.429	13.816
3	0.0717	0.216	4.642	6.251	7.815	9.348	9.837	11.345	12.838	14.796	16.266
4	0.207	0.484	5.989	7.779	9.488	11.143	11.668	13.277	14.860	16.924	18.467
5	0.412	0.831	7.289	9.236	11.070	12.833	13.388	15.086	16.750	18.907	20.515
6	0.676	1.237	8.558	10.645	12.592	14.449	15.033	16.812	18.548	20.791	22.458
7	0.989	1.690	9.803	12.017	14.067	16.013	16.622	18.475	20.278	22.601	24.322
8	1.344	2.180	11.030	13.362	15.507	17.535	18.168	20.090	21.955	24.352	26.124
9	1.735	2.700	12.242	14.684	16.919	19.023	19.679	21.666	23.589	26.056	27.877
10	2.156	3.247	13.442	15.987	18.307	20.483	21.161	23.209	25.188	27.722	29.588
11	2.603	3.816	14.631	17.275	19.675	21.920	22.618	24.725	26.757	29.354	31.264
12	3.074	4.404	15.812	18.549	21.026	23.337	24.054	26.217	28.300	30.957	32.909
13	3.565	5.009	16.985	19.812	22.362	24.736	25.472	27.688	29.819	32.535	34.528
14	4.075	5.629	18.151	21.064	23.685	26.119	26.873	29.141	31.319	34.091	36.123
15	4.601	6.262	19.311	22.307	24.996	27.488	28.259	30.578	32.801	35.628	37.697
16	5.142	6.908	20.465	23.542	26.296	28.845	29.633	32.000	34.267	37.146	39.252
17	5.697	7.564	21.615	24.769	27.587	30.191	30.995	33.409	35.718	38.648	40.790
18	6.265	8.231	22.760	25.989	28.869	31.526	32.346	34.805	37.156	40.136	42.312
19	6.844	8.907	23.900	27.204	30.144	32.852	33.687	36.191	38.582	41.610	43.820
20	7.434	9.591	25.038	28.412	31.410	34.170	35.020	37.566	39.997	43.072	45.315
21	8.034	10.283	26.171	29.615	32.671	35.479	36.343	38.932	41.401	44.522	46.797
22	8.643	10.982	27.301	30.813	33.924	36.781	37.659	40.289	42.796	45.962	48.268
23	9.260	11.689	28.429	32.007	35.172	38.076	38.968	41.638	44.181	47.391	49.728
24	9.886	12.401	29.553	33.196	36.415	39.364	40.270	42.980	45.559	48.812	51.179
25	10.520	13.120	30.675	34.382	37.652	40.646	41.566	44.314	46.928	50.223	52.620
26	11.160	13.844	31.795	35.563	38.885	41.923	42.856	45.642	48.290	51.627	54.052
27	11.808	14.573	32.912	36.741	40.113	43.195	44.140	46.963	49.645	53.023	55.476
28	12.461	15.308	34.027	37.916	41.337	44.461	45.419	48.278	50.993	54.411	56.892
29	13.121	16.047	35.139	39.087	42.557	45.722	46.693	49.588	52.336	55.792	58.301
30	13.787	16.791	36.250	40.256	43.773	46.979	47.962	50.892	53.672	57.167	59.703

Figure 1

Looking in the "5" row, we see that 1.6 is way below even the 0.2 significance threshold (if we are sensible we would typically use something lower like 0.05 or 0.01), so we do not reject the idea that the dice is fair. The do not reject language is weird, I know, but statisticians use it so I will follow the convention.

Now here is another rule that feels like magic, but actually there isn't much to prove here since it's based on numerical data: If cells have expected counts under 5 we combine them (since the chi squared statistic is less accurate and n needs to be made larger). Here is an example.

Lets say that we want to test whether $X \sim Po(2.5)$ and we roll 120 instances of X. Then the table of expected counts is as follows (we could find it using poisson distribution tables or calculating the probabilities using the formula directly):

0	1	2	3	4	5	6	≥ 7
~9.85	~24.62	~30.78	~25.65	~16.03	~8.02	~3.34	~1.70

We see that the last 2 cells are each under 5. Therefore, we combine them. Our new expected counts are as follows:

0	1	2	3	4	5	≥ 6
~9.85	~24.62	~30.78	~25.65	~16.03	~8.02	~5.04

These are all legal.

Lets do a test with this just to practice, at the 0.05 significance level. It's fun to do this when I actually know what it takes to prove it, lol.

0	1	2	3	4	5	≥ 6
4	20	38	19	18	12	9

There are 6 degrees of freedom (7 cells minus 1 constraint which is the total), so the chi squared statistic (which turns out to be about 13.09) is above the 0.05 significance level, so we reject the null hypothesis.

Ok, here is yet another rule that feels like magic. Suppose we are given the same data as above and we want to test if it comes from a poisson distribution - only this time we *don't know* the mean and estimate it from the data. The same rule

works if we are testing a binomial and (with care - I will explain shortly) a normal distribution.

For the sake of example we will assume that all $9 \geq 6$ instances were exactly 6, ie our data is

0	1	2	3	4	5	6
4	20	38	19	18	12	9

Now the estimated mean is as follows:

$$\frac{0*4+1*20+2*38+3*19+4*18+5*12+6*9}{120} = \frac{339}{120} = 2.825$$

Therefore we ought to test for a $Po\,(2.825)$ distribution. But we are estimating a parameter now so we have to subtract another degree of freedom. This is madness, if you haven't studied the topic of how to prove this as much as I have it seems like I'm making up magic random nonsense rules and it's so frustrating - this is exactly what inspired me to make levels 4 and 6 started as an A level proof companion in the first place. How the hell is estimating a parameter the same as a constraint - we will see a nice explanation in level 6.3. We will see in level 6.3 what kinds of parameters actually follow this rule (It is not all of them, since there are one-to-one mappings between the set of real numbers and the set of pairs of real numbers so we have to have some regularity. We will show that in particular it works for "linear combinations" of the cell counts like above, as long as we satisfy some dimension constraints, see level 6.3 for more details.)

Lets do the test, now noting that the rightmost cell is ≥ 6 instead of just 6 and our estimated mean is 2.825. We test against a Po(2.825) distribution. We get a value of 7.16 as our statistic. Comparing to the 5 degrees of freedom entry of the table this is not statistically significant. Therefore we conclude that Poisson distribution is a suitable model for the data.

Now, here is an example of testing a normal distribution. We cannot estimate the variance here - since it does not satisfy the conditions that we prove in Level 6.3 and THANK GOODNESS I have never seen an A level question on this that would force me to prove it despite the fact that it is technically in some specifications so if you see one for my sanity PLEASE DO NOT TELL ME as I would have to deal with the bias issue and the nonlinearity and all that nonsense and it would not be appropriate for level 6 - but we can estimate the mean and fix the variance. Lets say people have some heights and we want to test if they follow a normal distribution with mean 170cm and variance 8cm. We will have some data in a table as follows:

Height (cm)	<155	155-160	160-165	165-170	170-175	175-180	180-185	>185
Frequency	9	14	24	33	45	43	19	13

Then we will find the probabilities, based on the normal model. To estimate the mean, we would assume all data is at the midpoint, for example, we would do something like

$$\frac{4*152.5+7*157.5+11*162.5+etc}{200}$$

Anyway lets just do the test.

We get a statistic of about 20.28 which is statistically significant for 7 degrees of freedom. This suggests that either my data is biased or my mean and standard deviation are off.

Finally we have the contingency table case. This, and the parameter estimation case above, are *particularly* annoying to prove, since many texts only prove the simple case. This is how I ended up obsessing over the problem so much.

Anyway, here is how the contingency table case works. Lets say, for example, we want to test if there is correlation between peoples Math grades and English grades and the data is as follows: (Sorry I can't figure out how to fix the table format someone let me know if you know how to)

		English		
		A	B	C
Math	A	18	27	19
	B	37	45	15
	C	13	11	15

Then if they are independent, we would have

$$P\left(\text{A in math and A in english}\right) = P\left(\text{A in math}\right) P\left(\text{A in english}\right)$$

The proportion is the number divided by the total, so we would have

$$\frac{E_{A,A}}{\text{Total}} = \frac{E_{\text{Total row 1}}}{\text{Total}} \frac{E_{\text{Total column 1}}}{\text{Total}}$$

Multiplying both sides by total, and generalizing, we get

$$E_{\text{Row a, Column b}} = \frac{E_{\text{Total row 1}} * E_{\text{Total column 1}}}{\text{Total}}$$

We will, for convenience, fill in the totals in the table.

		English			Total
		A	B	C	
Math	A	18	27	19	64
	B	37	45	15	97
	C	13	11	15	39
Total		68	83	49	200

Therefore, using the formula above, we get the following expected counts:

		English			Total
		A	B	C	
Math	A	21.76	26.56	15.68	64
	B	32.98	40.255	23.765	97
	C	13.26	16.185	9.555	39
Total		68	83	49	200

Now, let r be the number of rows and c be the number of columns. Now it may not be surprising that there is another "magic" rule. The number of degrees of freedom is $(r-1)(c-1)$ which is 4 in this example. We can calculate the sum $\sum \frac{(O-E)^2}{E}$ and compare it against the χ^2_4 distribution (By the way, that is the notation for a chi squared distribution with 4 degrees of freedom). The statistic is about 10.41 which is above the 0.05 significance level for 4 degrees of freedom, so we conclude that there is a correlation.

We will see in level 6.3 that the contingency table case and the parameter estimation case are very similar problems in a sense.

8 Hypothesis tests with the normal distribution

8.1 Testing difference between means of normals with known variance

Setup: Let X and Y be normal with means μ_X and μ_Y and standard deviations σ_X and σ_Y respectively. Let X be rolled n_X times and Y be rolled n_Y times.

Now the distribution of $\bar{X} - \bar{Y}$ has mean $\mu_X - \mu_Y$ and variance $Var\left(\bar{X}\right) + Var\left(\bar{Y}\right) = \frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}$.

Now, it follows that the variable

$$Z := \frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}}$$

is an $N(0,1)$, since we took a normal, subtracted its mean, and divided by the standard deviation.

Therefore, if $\mu_X = \mu_Y$ then $\frac{\left(\bar{X} - \bar{Y}\right)}{\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}}$ will be a standard normal which we can test against using a table.

Example. Suppose that in a sample of 40 boys and 30 girls, the mean weight of the boys was 58kg and the mean weight of the girls was 55kg and the standard deviations are known to be 5kg for the boys and 8kg for the girls. Then we test if there is a difference towards boys by calculating the statistic above. Let X represent girls and Y represent boys. Lets say that we have a 5% significance level. In this case,

$$\frac{\left(\bar{X}-\bar{Y}\right)}{\sqrt{\frac{\sigma_X^2}{n_X}+\frac{\sigma_Y^2}{n_Y}}}=\frac{-3}{\sqrt{\frac{64}{30}+\frac{25}{40}}}=\frac{-3}{\sqrt{\frac{64}{30}+\frac{25}{40}}}\approx -1.806$$

Since the 5% critical values are about ±1.64 from a table we conclude that there is evidence of a difference.

8.2 Testing the variance of a normal distribution

We will start by stating something which we will not prove here, but we will prove it in level 6.3.

If we have a random sample of n observations $X_1, X_2, ..., X_n$ from a normal distribution with unknown mean and variance σ^2 then the unbiased estimator for the variance S^2 will be such that $\frac{(n-1)S^2}{\sigma^2}$ follows a chi squared distribution with n-1 "degrees of freedom". I use quotation marks because the idea of calling it that seems so dodgey to me.

Here is how we can use this to test the variance of a normal distribution. We use the chi squared table from above and we consider both the lower and upper percentiles. Then with 95% probability we will have

$$\chi^2_{n-1}\left(0.975\right)<\frac{\left(n-1\right)S^2}{\sigma^2}<\chi^2_{n-1}\left(0.025\right)$$

Rearranging gives

$$\frac{\left(n-1\right)S^2}{\chi^2_{n-1}\left(0.025\right)}<\sigma^2<\frac{\left(n-1\right)S^2}{\chi^2_{n-1}\left(0.975\right)}$$

This gives us a way given our n observations to use a chi squared table to get our critical values.

Example. Suppose our data is 21, 25, 23, 19, 17, 22, 29, 25. Then $n = 8$ so the number of "degrees of freedom" is 7. Here the mean is 22.625 and s^2 is about 14.27 (we can use a calculator to find these). Using the table in the figure above we get that $\chi^2_7\left(0.025\right)$ is about 16.01 and $\chi^2_7\left(0.975\right)$ is about 1.69. Now we can use the formula $\frac{(n-1)S^2}{\chi^2_{n-1}(0.025)}<\sigma^2<\frac{(n-1)S^2}{\chi^2_{n-1}(0.975)}$ and we get a 95% confidence interval for the variance as being between 6.24 and 59.10 (to 2 decimal places). This seems like quite a sparse interval, but if we square root it then it is from about 2.50 to 7.69 for the standard deviation, which is still more than a factor of 3 but it makes sense since we don't have that much data.

If we are given $\sum x$ and $\sum x^2$ then recall that

$$\sum\left(x-\bar{x}\right)^2=\sum x^2-2\bar{x}\sum x+n\bar{x}^2=\sum x^2-n\bar{x}^2=\sum x^2-\frac{\left(\sum x\right)^2}{n}$$

so s^2 which is $\frac{1}{n-1}\sum\left(x-\bar{x}\right)^2$ would be given by the formula

$$\frac{1}{n-1}\left(\sum x^2-\frac{\left(\sum x\right)^2}{n}\right)$$

8.3 Testing if two normal distributions have the same variance

Lets say X and Y are 2 normal distributions with unknown mean and variance. Then we know from section 8.2 that

$$\frac{\left(n_x-1\right)S_x^2}{\sigma_x^2}\sim\chi^2_{n_x-1}\text{ and }\frac{\left(n_y-1\right)S_y^2}{\sigma_y^2}\sim\chi^2_{n_y-1}$$

Rearranging gives

$$\frac{\left(\frac{S_x^2}{\sigma_x^2}\right)}{\left(\frac{S_y^2}{\sigma_y^2}\right)}\sim\frac{\left(\frac{\chi^2_{n_x-1}}{n_x-1}\right)}{\left(\frac{\chi^2_{n_y-1}}{n_y-1}\right)}$$

The distribution above is called an F distribution. The F distribution has 2 parameters, in this case the two parameters would be $n_x - 1$ and $n_y - 1$. In general an $F(a, b)$ distribution is equal to

$$\frac{\left(\frac{\chi_a^2}{a}\right)}{\left(\frac{\chi_b^2}{b}\right)}$$

Note that if $\sigma_x = \sigma_y$ then $\frac{\left(\frac{S_x^2}{\sigma_x^2}\right)}{\left(\frac{S_y^2}{\sigma_y^2}\right)} = \frac{\left(S_x^2\right)}{\left(S_y^2\right)}$ is what we are testing, and we are testing it against an $F\left(n_x - 1, n_y - 1\right)$ distribution. The table of 5% critical values for this distribution is shown in Figure 2. Note that the way it works is that we would look in the $n_x - 1$'th column and the $n_y - 1$'th row to find the correct critical value.

	1	2	3	4	5	6	7	8	9	10
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16

Figure 2

These are the upper critical values. But the F distribution may also be too low which would suggest that Y has a larger variance than X. If we are only testing if X has a larger variance, this is a one tailed test so we just use the upper critical value, otherwise we use the lower one. To find the lower critical value, note that

- The upper critical value of $\frac{\left(S_y^2\right)}{\left(S_x^2\right)}$ is 1 divided by the lower critical value of $\frac{\left(S_x^2\right)}{\left(S_y^2\right)}$.

Therefore, the 5% lower critical value for (a, b) in the table is the reciprocal of the upper critical value in (a, b) which we can find from the table.

Example. Suppose 2 samples with size 6 and 7 have sample variances 13 and 18 respectively, and we want to do a two tailed test on it. Then the upper critical value at $(5, 6)$ in the F table above is 4.39, which is much bigger than $\frac{13}{18}$ so we are not critical in that direction. The reciprocal of the $(6, 5)$ value is the lower critical value and it is about 0.202 which is way below $\frac{13}{18}$ so the conclusion is we do not reject the null hypothesis.

8.4 Testing the mean of a normal distribution with unknown variance

We can test for means if the variance is known - We found some confidence intervals earlier in this document. However, if the variance is not known.

As we saw previously, if the sample size is small, $s = \sqrt{S^2}$ may not be very close to σ and so $\frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}}$ may not be close to an $N\left(0, 1\right)$ distribution. However, the above quantity does follow a distribution called the t distribution with parameter $n - 1$. As the sample size gets large, the uncertainty in the standard deviation diminishes and the t distribution does appraoch a standard normal.

Although this is beyond the scope of this level, s turns out NOT to be an unbiased estimator for σ.

It is easy to see that the t distribution does not depend on the variance - it is clear if we fix the variance to 1, and it stays the same when we rescale it.

Therefore, given a null hypothesis for μ we can compute $\frac{\bar{X}-\mu}{\frac{s}{\sqrt{n}}}$ and compare it against a table of the t distribution with parameter n-1 (We call this the number of "degrees of freedom" and as I am writing this I have no idea why like what even is the "constraint" here it makes no sense). But you hopefully get the idea of comparing things against tables by now. Figure 3 shows a table for the t distribution.

	P						
one-tail	0.1	0.05	0.025	0.01	0.005	0.001	0.0005
two-tails	0.2	0.1	0.05	0.02	0.01	0.002	0.001
DF							
1	3.078	6.314	12.706	31.821	63.656	318.289	636.578
2	1.886	2.92	4.303	6.965	9.925	22.328	31.6
3	1.638	2.353	3.182	4.541	5.841	10.214	12.924
4	1.533	2.132	2.776	3.747	4.604	7.173	8.61
5	1.476	2.015	2.571	3.365	4.032	5.894	6.869
6	1.44	1.943	2.447	3.143	3.707	5.208	5.959
7	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	1.397	1.86	2.306	2.896	3.355	4.501	5.041
9	1.383	1.833	2.262	2.821	3.25	4.297	4.781
10	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	1.356	1.782	2.179	2.681	3.055	3.93	4.318
13	1.35	1.771	2.16	2.65	3.012	3.852	4.221
14	1.345	1.761	2.145	2.624	2.977	3.787	4.14
15	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	1.337	1.746	2.12	2.583	2.921	3.686	4.015
17	1.333	1.74	2.11	2.567	2.898	3.646	3.965
18	1.33	1.734	2.101	2.552	2.878	3.61	3.922
19	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	1.325	1.725	2.086	2.528	2.845	3.552	3.85
21	1.323	1.721	2.08	2.518	2.831	3.527	3.819
22	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	1.319	1.714	2.069	2.5	2.807	3.485	3.768
24	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	1.316	1.708	2.06	2.485	2.787	3.45	3.725
26	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	1.314	1.703	2.052	2.473	2.771	3.421	3.689
28	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	1.311	1.699	2.045	2.462	2.756	3.396	3.66
30	1.31	1.697	2.042	2.457	2.75	3.385	3.646

Figure 3

We can also do something called a paired t test. The procedure is as follows. Suppose that we have 2 sets of normally distributed data (or we assume they are normal). Suppose they have size n and we have $X_1, X_2, ..., X_n$ and $Y_1, Y_2, ..., Y_n$, each X_i related somehow to the Y_i. For instance, it might be that X_i is the parameter for someone before some treatment and Y_i is the parameter after and we want to test if there is a difference. In this case, we just apply the normal t test above to $X_i - Y_i$ and test if it has mean 0 against a t_{n-1} table.

Example. Suppose we want to test if being drunk increases reaction times and the data is as follows.

Before (X)	0.4	0.6	0.4	0.2	0.6	0.8	0.9	0.4	1.0	0.8
After (Y)	0.6	0.8	0.7	0.5	0.6	0.9	0.7	0.8	1.0	0.7
Y-X	0.2	0.2	0.3	0.3	0.0	0.1	0.2	0.4	0.0	-0.1

Here, s^2 for the bottom row is about 0.037, so t, which under the null hypothesis $\mu = 0$ is $\frac{\bar{X}}{\sqrt{\frac{s^2}{n}}}$ which is $\frac{0.12}{\sqrt{\frac{0.037}{10}}}$ which is about 1.96. Since the 5% critical value in the 10-1=9 row in the table was 1.833, it is significant and we conclude there is evidence of an increase in reaction time.

Note that we are testing for a positive t value if we want to know if the mean is greater than the null hypothesis mean, a negative one if we want to know if the mean is less, and an extreme one either way for a two tailed test.

8.5 Testing difference between means of normals with equal but unknown variance

To do what we have suggested in the title of this subsection, we calculate a pooled estimate of the variance as follows:

$$\frac{n_x - 1}{n_x + n_y - 2} s_x^2 + \frac{n_y - 1}{n_x + n_y - 2} s_y^2 = \frac{(n_x - 1) s_x^2 + (n_y - 1) s_y^2}{n_x + n_y - 2}$$

Where recall $s_x^2 = \frac{1}{n_x - 1} \sum (x - \bar{x})^2$ and similarly for s_y^2. We will call this "pooled estimate" s_p^2. Now, something we will prove in level 6.3 is that

$$\frac{\left(\bar{X} - \bar{Y}\right) - (\mu_X - \mu_Y)}{S_p \sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}} \sim t_{n_X + n_Y - 2}$$

So we want to test $\frac{\left(\bar{X} - \bar{Y}\right)}{S_p \sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}}$ (since we are assuming the other term is 0) against a $t_{n_X + n_Y - 2}$ distribution.

I've done enough examples, you get the idea.