Topic 20: Hypothesis Testing

02-680: Essentials of Mathematics and Statistics

November 26, 2024

Once we have taken our data and tried to fit a model to it, we often want to know if that fit model matches our original assumptions about the data. We call this ***hypothesis testing***, and we use these techniques to put an actual value on this match.

We add some terminology on what we've been talking about with respect to statistics and frame things as follows: we first define our ***alternate hypothesis***, which we will denote H_1, this will be the set of events that define what we want to ask about the confidence in happening; we then define the ***null hypothesis***, which we denote H_0, this is the set of events that are all outcomes other than than our alternate. So lets say were asking if a drug has a measurable impact on cholesterol, the null hypothesis would be that cholesterol stayed the same and the alternate hypothesis would be that it changed.

We usually refer to a hypothesis test telling us if we should ***reject*** or ***retain*** the null hypothesis.

1 Defining Errors

Errors occur when the hypothesis test tells us something thats wrong. So in the example above if the test tells us to reject the null (that is, its confident that the cholesterol changed) but in reality it didn't change we call this a ***Type I*** error. On the other hand if our test tells us to retain the null but in reality the value *did* change we call that a ***Type II*** error.

		Hypothesis Test Result	
		Retain H_0	Reject H_0
Truth	H_0		Type I Error
	H_1	Type II Error	

We say the

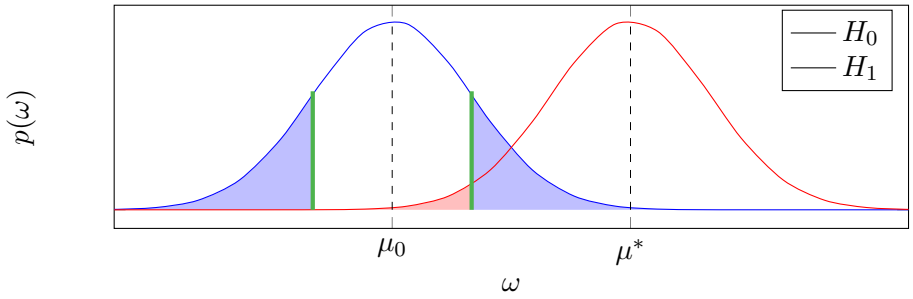
Type I Error Rate is $p(\text{reject } H_0, H_0 \text{ is true})$,

Type II Error Rate is $p(\text{retain } H_0, H_1 \text{ is true})$, and

Statistical ***Power*** is 1 - Type II Error Rate.

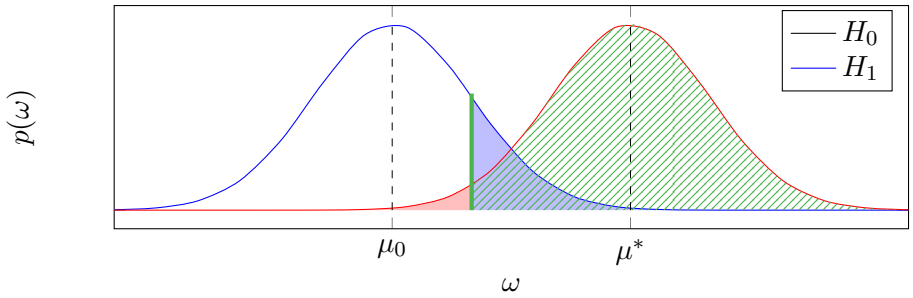
The last point means that the higher power tests have a stronger ability to detect signals for H_1.

Let's look at it visually, first for what we call a ***two-sided*** test, that is $H_0 : \mu = x$ and $H_1 : \mu \neq x$.



In the figure above, when we pick some boundary around our desired x (the green lines) we will have some probability of Type I Error (blue shaded regions) and Type II Error (red shaded region).

In a ***one-sided*** test we say $H_0 : \mu \leq x$ and $H_1 : \mu > x$.



In this example, one again the Type I and Type II errors are shown in the blue and red shaded regions, but we can also see the power of the test (green stripped region).

Main Takeaways. In both cases we can choose the cutoff (the thick green lines) of where to make the distinction between H_0 and H_1, but there is a tradeoff: **as Type I error goes down, Type II will go up** and power will go down.

Similarly, as μ^* and μ_0 become further apart both errors will go down, and the signal becomes easier to detect.

2 Performing Tests

Lets assume we have some data $\mathcal{D} = X_1, X_2, \cdots, X_n$ that follows the same distribution with known $p(X_i \mid \theta)$. We will perform a test in 3 steps:

(1) Compute a test statistic (a function of the data) thats appropriate for the distribution:

$$T = r(X_1, X_2, \cdots, X_n),$$

(2) Compute a p-value (we will talk about this below), then

(3) For a desired significance level β and the p-value, decide whether to retain or reject H_0.

We will use the ongoing example from above about testing the efficacy of a drug for high cholesterol, lets assume we know the typical variance of cholesterol among humans (σ) and that this is not going to change between the conditions.

Step 1. In this case, since we're assuming that the samples are coming from a Gaussian (Normal) distribution ($X_1, X_2, \cdots, X_n \sim \mathcal{N}(\mu_0, \sigma^2)$), the test statistic is the mean ($\overline{X_n}$). And lets assume we're running a two-sided test (we don't know if the impact is going to lower or raise cholesterol we just want to know if it changes); thus $H_0 : \mu = \mu_0$ and $H_1 : \mu \neq \mu_0$.

Under the null hypothesis

$$\overline{X_n} \sim \mathcal{N}(x, \sigma^2/n)$$

or as we saw

$$\frac{\overline{X_n} - \mu_0}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1).$$

Step 2. We want to compute the p-value, which is essentially the probability that we would see the test statistic under the null hypothesis (for T):

$$p\left(|T| > \frac{\overline{X_n} - \mu_0}{\sigma/\sqrt{n}}\right)$$

Step 3. Depending on how strict we want to be, we accept or reject H_0 based on the p-value. Some general rules on how to chose α:

p-value	interpretation
< 0.01	very strong evidence against H_0
$0.01 - 0.05$	strong evidence against H_0
$0.05 - 0.1$	weak evidence against H_0
> 0.1	little to no evidence against H_0

One-sided tests. For a one-sided test; that is one where for example $H_0 : \mu \leq \mu_0$ and $H_1 : \mu > \mu_0$; the only thing that changes is that in Step 2:

$$p\left(T > \frac{\overline{X_n} - \mu_0}{\sigma/\sqrt{n}}\right).$$

Notice that usually one-sided tests can be more powerful as they are only integrating over one region.

Main Takeaways. When performing tests, the main thing we need to do is determine:

- the distribution we think the data came from,
- the test statistic thats appropriate for those samples, and
- the distribution that applies to the test statistic
  (it may be different than the one for the data).

3 t-tests: When σ is Unknown

Lets again assume $X_1, X_2, \cdots, X_n \sim \mathcal{N}(\mu_0, \sigma^2)$ but this time we don't know σ. We're going to define two statistics:

$$\overline{X_n} = \frac{1}{n}\sum X_i$$

and

$$\overline{\sigma}^2 = \frac{1}{n-1}\sum (X_i - \overline{X_n})^2$$

We're then going to say the following:

$$\frac{\overline{X_n} - \mu_0}{\overline{\sigma}/\sqrt{n}} \sim t_{n-1}$$

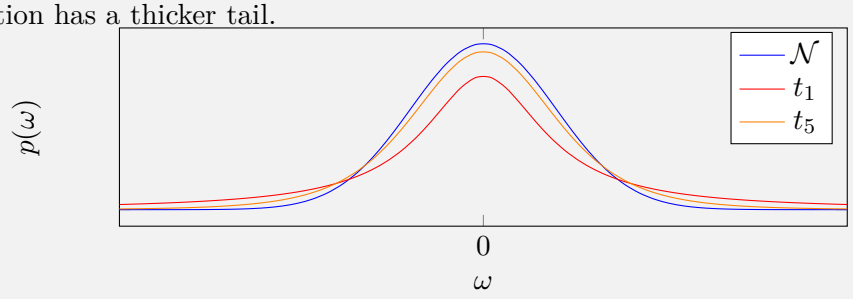
The t distribution

The t distribution has one parameter: ν, which is the number of *degrees of freedom*.

$$X \sim t_\nu$$

$$p(X = x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi\nu} \cdot \Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

The shape of t distribution is similar to the shape of normal distribution but t distribution has a thicker tail.



As $v \to \infty$, the distribution above becomes a normal distribution.

So we have a way to compute something that should be in t_{n-1}, that is, we assume there is one less degree of freedom than there are elements in the observation.

Since the probability of the t distribution is difficult to calculate, we typically have lookup tables for it (see below). To use one of these tables find your degrees of freedom in the left column and use that row to find the column with next smaller number from your statistic. Read the probability in the top row. Since your t will probably be a little bit bigger than the value in the table, your p will be smaller, eg., $p < 0.01$. If your t is to the right of all numbers, then p is less than the right most probability.

4 Paired Data: Paired t-tests

Many times the data we have is a set of samples measured in two different conditions. As an example, a set of patients measured before an after treatment. In that case we want to know if the treatment made a consistent change across the population. We also don't know where each of the individuals sits compared with some (unknown) μ.

In these cases our hypotheses are:

$$H_0 : X_1 = X_2 \quad \text{and} \quad H_1 : X_1 \neq X_2$$

(where X_1 and X_2 are the two experimental conditions). Another way to say this is to

define some $\delta = X_1 - X_2$, and let the hypotheses be

$$H_0 : \delta = 0 \quad \text{and} \quad H_1 : \delta \neq 0.$$

As we did before we can compute $\overline{\delta_n}$ and $\overline{\sigma}_\delta$ from the data. and let our statistic be

$$\frac{\overline{\delta_n} - \mu_0}{\overline{\sigma_\delta}/\sqrt{n}} \sim t_{n-1}$$

5 Testing Categorical Data: χ^2 tests

Sometimes we have data where we have some underlying conceptual probabilities for a group of categories, and want to know how well what we observed fits this concept.

Specifically, assume we have some underlying assumption that we will see a set of n categories with the following probabilities: $\dot{p} = (\dot{p}_1, \dot{p}_2, \cdots, \dot{p}_n)$. And a set of observations, which we converted to probabilities $p = (p_1, p_2, \cdots, p_n)$. We then want to test

$$H_0 : \dot{p} = p \quad \text{and} \quad H_1 : \dot{p} \neq p$$

(in both cases we assume $\sum \dot{p}_i = 1$ and $\sum p_i = 1$.)

A good example of this is Mendel's pea experiment. Mendel's hypothesis was that the proportion of round/yellow peas, wrinkled/yellow peas, round/green peas, and wrinkled/green peas is given as

$$\dot{p} = \left(\frac{9}{16}, \frac{3}{16}, \frac{3}{16}, \frac{1}{16}\right)$$

Lets assume the observations were as follows:

	round/yellow	wrinkled/yellow	round/green	wrinkled/green	total
count	315	101	108	32	556
expected counts	312.75	104.25	104.25	34.75	556

In this case the test statistic will be

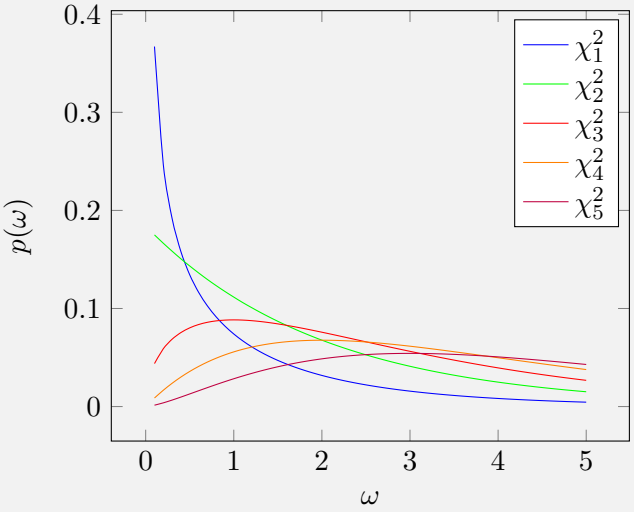
$$\sum \frac{(c_i - n \cdot \dot{p}_i)^2}{\dot{p}_i)} \sim \chi^2_{n-1}$$

The χ^2 distribution

$$X \sim \chi^2_p$$

$$p(X = x) = \frac{x^{\frac{p}{2}-1}}{\Gamma\left(\frac{p}{2}\right) 2^{\frac{p}{2}}} e^{\frac{-x}{2}}$$

Here p is the number of degrees of freedom, and in all cases $x > 0$.



If $Z_1, Z_2, \cdots, Z_p$ are independent standard normal random variables, then

$$\sum Z_i^2 \sim \chi^2_p$$

Looking at the previous example with Mendel's pea plants.

$$\sum_{i=1} 4 \frac{(c_i - n \cdot \dot{p}_i)^2}{\dot{p}_i)}$$

$$= \frac{(315 - 312.75)^2}{312.75} + \frac{(101 - 104.25)^2}{104.25} + \frac{(108 - 104.25)^2}{104.25} + \frac{(32 - 34.75)^2}{34.75} \approx 0.47$$

Just like with t-tests we typically look up the p-value for χ^2 tests in a table. One is included below. So we have 3 degrees of freedom, thus we cannot reject the null hypothesis.

Useful References

Wasserman. "All of Statistics: A Concise Course in Statistical Inference" §10
Degroot and Schervish. "Probability and Statistics" §9

t Distribution Lookup

ν	0.4	0.33	0.25	0.2	0.125	0.1	0.05	0.025	0.01	0.005	0.001
1	0.325	0.577	1.000	1.376	2.414	3.078	6.314	12.706	31.821	63.657	318.31
2	0.289	0.500	0.816	1.061	1.604	1.886	2.920	4.303	6.965	9.925	22.327
3	0.277	0.476	0.765	0.978	1.423	1.638	2.353	3.182	4.541	5.841	10.215
4	0.271	0.464	0.741	0.941	1.344	1.533	2.132	2.776	3.747	4.604	7.173
5	0.267	0.457	0.727	0.920	1.301	1.476	2.015	2.571	3.365	4.032	5.893
6	0.265	0.453	0.718	0.906	1.273	1.440	1.943	2.447	3.143	3.707	5.208
7	0.263	0.449	0.711	0.896	1.254	1.415	1.895	2.365	2.998	3.499	4.785
8	0.262	0.447	0.706	0.889	1.240	1.397	1.860	2.306	2.896	3.355	4.501
9	0.261	0.445	0.703	0.883	1.230	1.383	1.833	2.262	2.821	3.250	4.297
10	0.260	0.444	0.700	0.879	1.221	1.372	1.812	2.228	2.764	3.169	4.144
11	0.260	0.443	0.697	0.876	1.214	1.363	1.796	2.201	2.718	3.106	4.025
12	0.259	0.442	0.695	0.873	1.209	1.356	1.782	2.179	2.681	3.055	3.930
13	0.259	0.441	0.694	0.870	1.204	1.350	1.771	2.160	2.650	3.012	3.852
14	0.258	0.440	0.692	0.868	1.200	1.345	1.761	2.145	2.624	2.977	3.787
15	0.258	0.439	0.691	0.866	1.197	1.341	1.753	2.131	2.602	2.947	3.733
16	0.258	0.439	0.690	0.865	1.194	1.337	1.746	2.120	2.583	2.921	3.686
17	0.257	0.438	0.689	0.863	1.191	1.333	1.740	2.110	2.567	2.898	3.646
18	0.257	0.438	0.688	0.862	1.189	1.330	1.734	2.101	2.552	2.878	3.610
19	0.257	0.438	0.688	0.861	1.187	1.328	1.729	2.093	2.539	2.861	3.579
20	0.257	0.437	0.687	0.860	1.185	1.325	1.725	2.086	2.528	2.845	3.552
21	0.257	0.437	0.686	0.859	1.183	1.323	1.721	2.080	2.518	2.831	3.527
22	0.256	0.437	0.686	0.858	1.182	1.321	1.717	2.074	2.508	2.819	3.505
23	0.256	0.436	0.685	0.858	1.180	1.319	1.714	2.069	2.500	2.807	3.485
24	0.256	0.436	0.685	0.857	1.179	1.318	1.711	2.064	2.492	2.797	3.467
25	0.256	0.436	0.684	0.856	1.178	1.316	1.708	2.060	2.485	2.787	3.450
26	0.256	0.436	0.684	0.856	1.177	1.315	1.706	2.056	2.479	2.779	3.435
27	0.256	0.435	0.684	0.855	1.176	1.314	1.703	2.052	2.473	2.771	3.421
28	0.256	0.435	0.683	0.855	1.175	1.313	1.701	2.048	2.467	2.763	3.408
29	0.256	0.435	0.683	0.854	1.174	1.311	1.699	2.045	2.462	2.756	3.396
30	0.256	0.435	0.683	0.854	1.173	1.310	1.697	2.042	2.457	2.750	3.385
35	0.255	0.434	0.682	0.852	1.170	1.306	1.690	2.030	2.438	2.724	3.340
40	0.255	0.434	0.681	0.851	1.167	1.303	1.684	2.021	2.423	2.704	3.307
45	0.255	0.434	0.680	0.850	1.165	1.301	1.679	2.014	2.412	2.690	3.281
50	0.255	0.433	0.679	0.849	1.164	1.299	1.676	2.009	2.403	2.678	3.261
55	0.255	0.433	0.679	0.848	1.163	1.297	1.673	2.004	2.396	2.668	3.245
60	0.254	0.433	0.679	0.848	1.162	1.296	1.671	2.000	2.390	2.660	3.232
∞	0.253	0.431	0.674	0.842	1.150	1.282	1.645	1.960	2.326	2.576	3.090

χ^2 $p$value lookup table

ν	0.4	0.33	0.25	0.2	0.125	0.1	0.05	0.025	0.01	0.005	0.001
1	0.708	0.936	1.323	1.642	2.354	2.706	3.841	5.024	6.635	7.879	10.828
2	1.833	2.197	2.773	3.219	4.159	4.605	5.991	7.378	9.210	10.597	13.816
3	2.946	3.405	4.108	4.642	5.739	6.251	7.815	9.348	11.345	12.838	16.266
4	4.045	4.579	5.385	5.989	7.214	7.779	9.488	11.143	13.277	14.860	18.467
5	5.132	5.730	6.626	7.289	8.625	9.236	11.070	12.833	15.086	16.750	20.515
6	6.211	6.867	7.841	8.558	9.992	10.645	12.592	14.449	16.812	18.548	22.458
7	7.283	7.992	9.037	9.803	11.326	12.017	14.067	16.013	18.475	20.278	24.322
8	8.351	9.107	10.219	11.030	12.636	13.362	15.507	17.535	20.090	21.955	26.125
9	9.414	10.215	11.389	12.242	13.926	14.684	16.919	19.023	21.666	23.589	27.877
10	10.473	11.317	12.549	13.442	15.198	15.987	18.307	20.483	23.209	25.188	29.588
11	11.530	12.414	13.701	14.631	16.457	17.275	19.675	21.920	24.725	26.757	31.264
12	12.584	13.506	14.845	15.812	17.703	18.549	21.026	23.337	26.217	28.300	32.910
13	13.636	14.595	15.984	16.985	18.939	19.812	22.362	24.736	27.688	29.819	34.528
14	14.685	15.680	17.117	18.151	20.166	21.064	23.685	26.119	29.141	31.319	36.123
15	15.733	16.761	18.245	19.311	21.384	22.307	24.996	27.488	30.578	32.801	37.697
16	16.780	17.840	19.369	20.465	22.595	23.542	26.296	28.845	32.000	34.267	39.252
17	17.824	18.917	20.489	21.615	23.799	24.769	27.587	30.191	33.409	35.718	40.790
18	18.868	19.991	21.605	22.760	24.997	25.989	28.869	31.526	34.805	37.156	42.312
19	19.910	21.063	22.718	23.900	26.189	27.204	30.144	32.852	36.191	38.582	43.820
20	20.951	22.133	23.828	25.038	27.376	28.412	31.410	34.170	37.566	39.997	45.315
21	21.991	23.201	24.935	26.171	28.559	29.615	32.671	35.479	38.932	41.401	46.797
22	23.031	24.268	26.039	27.301	29.737	30.813	33.924	36.781	40.289	42.796	48.268
23	24.069	25.333	27.141	28.429	30.911	32.007	35.172	38.076	41.638	44.181	49.728
24	25.106	26.397	28.241	29.553	32.081	33.196	36.415	39.364	42.980	45.559	51.179
25	26.143	27.459	29.339	30.675	33.247	34.382	37.652	40.646	44.314	46.928	52.620
26	27.179	28.520	30.435	31.795	34.410	35.563	38.885	41.923	45.642	48.290	54.052
27	28.214	29.580	31.528	32.912	35.570	36.741	40.113	43.195	46.963	49.645	55.476
28	29.249	30.639	32.620	34.027	36.727	37.916	41.337	44.461	48.278	50.993	56.892
29	30.283	31.697	33.711	35.139	37.881	39.087	42.557	45.722	49.588	52.336	58.301
30	31.316	32.754	34.800	36.250	39.033	40.256	43.773	46.979	50.892	53.672	59.703
35	36.475	38.024	40.223	41.778	44.753	46.059	49.802	53.203	57.342	60.275	66.619
40	41.622	43.275	45.616	47.269	50.424	51.805	55.758	59.342	63.691	66.766	73.402
45	46.761	48.510	50.985	52.729	56.052	57.505	61.656	65.410	69.957	73.166	80.077
50	51.892	53.733	56.334	58.164	61.647	63.167	67.505	71.420	76.154	79.490	86.661
55	57.016	58.945	61.665	63.577	67.211	68.796	73.311	77.380	82.292	85.749	93.168
60	62.135	64.147	66.981	68.972	72.751	74.397	79.082	83.298	88.379	91.952	99.607