



Hypothesis testing



Data to decisions

The idea

- Null hypothesis:

H_0 : the DGP/population has property P

- Under the null, a sample statistic has a known distribution
- If, under that that distribution, the value of the statistic is unlikely, reject the null



Road map

- This chapter illustrates the general procedure by discussing some of the most common tests people perform:
 1. Simple mean tests
 2. Mean comparison tests
 3. Frequency (or proportion) tests
 4. Goodness of fit tests
 5. Independence tests
- The next chapter applies the same procedure to the regression context



Simple mean test

- You believe that your customer base has mean income \$40,000
- A recent, representative survey of 1,000 customers showed their mean income to be \$37,000, with a standard deviation of \$2,000
- Is it time to revise your beliefs?



Mean test design

- $H_0: \mu = \$40,000$
- If we also knew σ (the population standard deviation) we would know that sample mean $\hat{\mu}$ is roughly normally distributed with mean \$40,000 and standard deviation $\frac{\sigma}{\sqrt{n}}$
- But we don't



The unknown sigma problem

- $\hat{\sigma} = 2,000$ is an estimate of σ
- It too is normally distributed by the CLT
- Test statistic:

$$T = \frac{\hat{\mu} - \mu}{\hat{\sigma} / \sqrt{n}}$$

- Under the null, this has a t-distribution with $n - 1$ degrees of freedom
 - Standard normal if n large
 - Reject, basically, if $T > 1.96$ or $T < -1.96$
 - Or look up t-tables for more precision.
-



Confidence intervals

- $\hat{\sigma}/\sqrt{n}$ is called the *standard error of the mean*
- For n large enough and **before we draw our data**, with 95% confidence we can say that the population mean μ should be in:

$$\left[\hat{\mu} - 1.96 \hat{\sigma}/\sqrt{n}, \hat{\mu} + 1.96 \hat{\sigma}/\sqrt{n} \right]$$

- This is called a *confidence interval* for the mean
 - If μ is outside this interval, reject the null with 95% confidence
-



General structure of t-tests

- When an estimate follows a normal distribution, then

$$\frac{\textit{Estimate} - \textit{null value}}{\textit{standard error of estimate}}$$

follows a t-distribution

- Only degrees of freedom need to be established and that is test-specific
- But, any time you have a large, representative sample, the distribution is approximately the standard normal so you are good to go



Remark

- In our current example, the odds that, literally, $\mu = \$40,000$, are, literally, zero
- Failing to reject that hypothesis means, simply, that that guess cannot be dismissed in favor of distant alternatives
- For instance, if you claim that $COV(X, Y)$ is positive and large, then you should be able to reject the hypothesis that $COV(X, Y) = 0$
- That's putting your theory to a test it should pass with flying colors



Critical values

- We can design a test by choosing a *significance level* (or *size* or *alpha*)
- Say we set $\alpha = 5\%$
- Then we can pick a critical value \bar{T} such that $P(T \geq \bar{T}) \leq 5\%$ if the null hypothesis is correct
- Reject if $T > \bar{T}$
- For normally distributed statistics: $\bar{T} = \mu + 1.655\sigma$
- Or we could pick two values \bar{T} and \underline{T} such that $P(T \geq \bar{T} \text{ or } T \leq \underline{T}) \leq 5\%$
- Reject if $T > \bar{T}$ or $T < \underline{T}$
- For normally distributed statistics, e.g.: $\bar{T} = \mu + 1.96\sigma, \underline{T} = \mu - 1.96\sigma$



Critical values vs p-values

- *p – values* look at the outcome of the test and then calculate its probability in some sense or other
- For instance, in the context of one-sided tests, a particular sample gives you a statistic value of \hat{T}
- The p-value is $P(T \geq \hat{T})$
- Ideally, you should design a test fully ex-ante (choose its size, in particular) and then let the data speak



Type 1 errors

- There is a risk that we may reject a null hypothesis when it is, in fact, correct
- When we use a 5% level to compute critical values, we create a test that has a 5% chance of producing a type I error
- This is often termed a “false positive” since rejecting H_0 is often viewed as “finding an effect.”



Type 2 errors

- There is a risk that we may fail to reject a null hypothesis when it is, in fact, incorrect
- The problem with this language is that since null hypotheses are usually quite specific, incorrect can mean a whole lot of different things
- It also means that H_0 taken literally, is often false (see remark slide)
- So how do people measure the risk of type 2 errors in practice?
- Answer: in a massively ad-hoc way
- For instance, in the context of one-sided tests for means with critical value \bar{T} , the risk of type 2 error is typically computed as the risk of getting a rejection when the truth is at \bar{T}



Mean comparison

- A university wants to know if it has a gender wage-gap problem
- It obtains a sample of male and female employees with similar education, age and occupation
- n_1 females, n_2 males
- Mean income among women \$97,000, stdev is \$1,000
- Mean income among men \$100,000, stdev is \$1,500
- $H_0: \mu_1 = \mu_2$
- Can it be rejected?



Test statistic

- $T = \frac{\hat{\mu}_2 - \hat{\mu}_1}{\sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}}}$ is t-distributed

- The expression for degrees of freedom looks nasty:

$$\frac{\left(\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}\right)^2}{\left(\frac{\hat{\sigma}_1^2}{n_1}\right)^2 / n_1 - 1 + \left(\frac{\hat{\sigma}_2^2}{n_2}\right)^2 / n_2 - 1}$$

- But that's why we have computers (Excel: ttest)
 - What's more, if $\hat{\sigma}_1 \approx \hat{\sigma}_2$, then degrees of freedom are roughly $n_1 + n_2 - 2$
 - And, if n_1 and n_2 are large, you can assume a standard normal distribution so, basically, reject $T > 1.96$ or $T < -1.96$
-



Frequency tests (the easiest of them all)

- H_0 : Machine true defect rate is $\pi = 1\%$
- The DGP/Population's standard deviation is $\sqrt{\pi(1 - \pi)}$
- So, for a sample of size n , the standard deviation (or *standard error*) of the mean is $\sqrt{\frac{\pi(1 - \pi)}{n}}$



Goodness of fit tests

- H_0 : Data came from process X
- Example: Are the 3,000 draws in data2D2D.xlsx from X ?
- If they were, I'd expect to see around 300 draws of 90, instead I see just 69 or them
- How far are the draws I got from what I'd expect under H_0 ?
- If they are farther than what sample uncertainty alone can reasonably explain, reject



Chisquare distance

- Data (O for “observed”, E for “expected”):

Value	Sample outcome (O)	Expected (E)	$(O-E)^2/E$
90	69	300	177.87
100	1,969	1,500	146.64
110	962	1,200	47.20
Total	3,000	3,000	371.71



Chisquare goodness of fit test

- If H_0 were correct, we'd expect the Chisquare distance to be small
- Specifically, if H_0 is correct, that distance roughly follows a Chisquare distribution
- Degrees of freedom = (number of values - 1) = 2 in this example
- A look at a Chisquare table says that 95% of the time a draw from such a distribution should be below 5.991
- We got a 371.71
- Reject H_0 with high confidence



Chisquare independence tests

- H_0 : Y is independent from X
- Ex: Is spending independent of gender in a particular population?
- If you have detailed data on spending from a good sample, you could run a regression
- But what if I only have coarse/categorical data?



Example

- Assume we get the following data from a representative sample:

Spending	Male (O)	Female (O)	All	Frequency
<50K	700	601	1,301	0.34
50-99	513	557	1,070	0.28
100-199	410	518	928	0.24
200 or more	227	309	536	0.14
Total	1,850	1,985	3,835	1



Expected outcome

- If Spending were independent of Gender, the frequency of observations in each spending group should be similar for both groups...
- ... hence the same as in the overall sample
- In other words, I'd expect something close to:

Spending	Male (E)	Female (E)
<50K	627.60	673.40
50-99	516.17	553.83
100-199	447.67	480.33
200 or more	258.57	277.43
Total	1,850	1,985



Chisquare distance

- The chisquare distance between observed and expected is:

Spending	Male $(O-E)^2/E$	Female $(O-E)^2/E$
<50K	8.35	7.78
50-99	0.02	0.02
100-199	3.17	2.95
200 or more	3.85	3.59
Total	15.39	14.34

- For a total distance of around 29.74
-



Chisquare independence tests

- H_0 : Spending is independent of gender
- If H_0 is correct, the chisquare distance roughly follows a chisquare distribution
- Degrees of freedom = (number of Spending categories - 1) times (number of Gender categories - 1) = 3
- A look at a Chisquare table shows that 95% of the time a draw from a Chisquare distribution with 3 degrees of freedom should be below 7.815
- We got 29.74, reject H_0 with confidence



The bootstrap

- Classical hypothesis testing relies on a lot of theory
- There is an alternative procedure that works for any statistic, however complicated, and requires few if any large sample assumptions: **bootstrapping**
- Idea:
 1. use available data to generate different samples hence a distribution of the statistic
 2. use that distribution to produce standard errors and/or produce confidence intervals
- One assumption: the sample is representative of the population
- An example will help illustrate the power of bootstrapping
- More generally, modern stats is putting ever more emphasis on methods which, like bootstrapping, require little to no theory

