Hypothesis Testing

A statement has been made. We must decide whether to believe it (or not). Our belief decision must ultimately stand on three legs:

- What does our general background knowledge and experience tell us (for example, what is the reputation of the speaker)?
- What is the cost of being wrong (believing a false statement, or disbelieving a true statement)?
- What does the relevant data tell us?

Making the "Belief" Decision

- What does our general background knowledge and experience tell us (for example, what is the reputation of the speaker)? – *The answer is typically already in the manager's head.*
- What is the cost of being wrong (believing a false statement, or disbelieving a true statement)? – *Again, the answer is typically already in the manager's head.*
- What does the relevant data tell us? – T*he answer is typically* **not** *originally in the manager's head. The goal of hypothesis testing is to put it there, in the simplest possible terms.*
- Then, the job of the manager is to pull these three answers together, and make the "belief" decision. **The statistical analysis contributes to this decision, but doesn't make it.**

Our Goal is Simple:

- To put into the manager's head a single phrase which summarizes all that the data says with respect to the original statement.

"The data, *all by itself,* makes me _____ suspicious, because the data, *all by itself,* contradicts the statement _____ strongly."
{not at all, a little bit, moderately, quite, very, overwhelmingly}

- We wish to choose the phrase which best fills the blanks.

What We *Won't* Do

- ~~Compute Pr(statement is true | we see this data).~~

(This depends on our prior beliefs, instead of just on the data. It requires that we pull those beliefs out of the manager's head.)

What We **Will** Do

- Compute Pr(we see this data | statement is true).

This depends *just* on the data. Since we don't expect to see improbable things on a regular basis, a small value makes us very suspicious.

This is Analogous to the British System of Criminal Justice

- The statement on trial – the so-called "null hypothesis" – is that "the accused is innocent."
- The prosecution presents evidence.
- The jury asks itself: "How likely is it that this evidence would have turned up, just by chance, if the accused really *is* innocent?"
- If this probability is close to 0, then the evidence strongly contradicts the initial presumption of innocence … and the jury finds the accused "Guilty!"

Example: Processing a Loan Application

- You're the commercial loan officer at a bank, in the process of reviewing a loan application recently filed by a local firm. Examining the firm's list of assets, you notice that the largest single item is $3 million in accounts receivable. You've heard enough scare stories, about loan applicants "manufacturing" receivables out of thin air, that it seems appropriate to check whether these receivables actually exist. You send a junior associate, Mary, out to meet with the firm's credit manager.
- Whatever report Mary brings back, your final decision of whether to believe that the claimed receivables exist will be influenced by the firm's general reputation, and by any personal knowledge you may have concerning the credit manager. As well, the consequences of being wrong – possibly approving an overly-risky loan if you decide to believe the receivables are there and they're not, or alienating a commercial client by disbelieving the claim and requiring an outside credit audit of the firm before continuing to process the application, when indeed the receivables are as claimed – will play a role in your eventual decision.

Processing a Loan Application

- Later in the day, Mary returns from her meeting. She reports that the credit manager told her there were 10,000 customers holding credit accounts with the firm. He justified the claimed value of receivables by telling her that the average amount due to be paid per account was at least $300.
- With his permission, she confirmed (by physical measurement) the existence of about 10,000 customer folders. (You decide to accept this part of the credit manager's claim.) She selected a random sample of 64 accounts at random, and contacted the account-holders. They all acknowledged soon-to-be-paid debts to the firm. The sample mean amount due was $280, with a sample standard deviation of $120.
- What do we make of this data? It contradicts the claim to some extent, but how strongly? It could be that the claim is true, and Mary simply came up with an underestimate due to the randomness of her sampling (her "exposure to sampling error").

Compute, then Interpret

- What do we make of Mary's data? We answer this question in two steps. First, we compute Pr(we see this data | statement is true). More precisely:

Pr (	conducting a study such as we just did, we'd see data at least as contradictory to the statement as the data we are, in fact, seeing	|	the statement is true, in a way that fits the observed data as well as possible	)

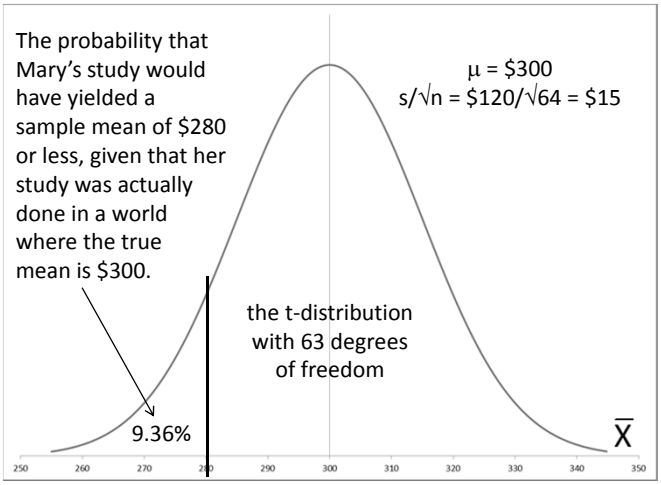
- This number is called ***the significance level of the data***, with respect to the statement under investigation (i.e., **with respect to *the null hypothesis***). (Some authors/software call this significance level the "p-value" of the data.)
- Then, we interpret the number: A small value forces us to say, "Either the statement is true, and we've been very unlucky (i.e., we've drawn a very misrepresentative sample), or the statement is false. We don't typically expect to be very unlucky, so the data, all by itself, makes us quite suspicious."

Null Hypothesis: "$\mu \geq \$300$"

- Mary's sample mean is \$280. Giving the original statement every possible chance of being found "innocent," we'll assume that Mary did her study in a world where the true mean is actually \$300.
- Let $\overline{X}$ be the estimate Mary might have gotten, had she done her study in this assumed world.

The significance level of Mary's data, with respect to the null hypothesis: "$\mu \geq \$300$", is

$$\Pr(\overline{X} \leq \$280 \mid \mu = \$300) = 9.36\%$$



The "Hypothesis Testing Tool"

- The spreadsheet "Hypothesis_Testing_Tool.xls," in the "Session 1" folder, does the required calculations automatically.

From the sample data, fill in the yellow boxes below:

280	estimate/prediction of unknown quantity
15	measure of uncertainty
64	sample size
0	number of explanatory variables in regression, or 0 if dealing with a population mean

Null hypothesis: true value ≥ = ≤ 300

significance level of data with respect to null hypothesis

9.3612%
18.7224%
100.0000%

(from t-distribution with 63 degrees of freedom)

And now, how do we interpret "9.36%"?

The "coin tossing" story.

numeric significance level of the data	interpretation: the data, all by itself, makes us	the data supports the alternative hypothesis
above 20%	not at all suspicious	not at all
between 10% and 20%	a little bit suspicious	a little bit
between 5% and 10%	moderately suspicious	moderately
between 2% and 5%	very suspicious	strongly
between 1% and 2%	extremely suspicious	very strongly
below 1%	overwhelmingly suspicious	incredibly strongly

9.36% is between 3 (12.5%) and 4 (6.25%) "tails" in a row. That's rare enough to make me "a bit" suspicious.

Processing a Loan Application

- So the data, all by itself, makes us "a bit suspicious." What do we do?
- It depends.
  - If the credit manager is a trusted lifelong friend …
  - If the credit manager is already under suspicion …
- What if Mary's sample mean were $260?
  - With a significance level of 0.49%, the data, all by itself, makes us "overwhelmingly suspicious" …

The One-Sided Complication

- A jury never finds the accused "innocent."
  - For example, if the prosecution presents no evidence at all, the jury simply finds the accused "not proven to be guilty."
- Just so, we never conclude that data ***supports*** the null hypothesis.
  - However, if data *contradicts* the null hypothesis, we *can* conclude that it supports the alternative.

So, If We Wish to Say that Data ***Supports*** a Claim …

- We take the opposite of the claim as our null hypothesis, and see if the data contradicts that opposite. If so, then we can say that the data supports the original claim.
- Examples:
  - A clinical test of a new drug will take as the null hypothesis that patients who take the drug are equally or less healthy than those who don't.
  - An evaluation of a new marketing campaign will take as the null hypothesis that the campaign is not effective.

From Session 1

- And now we return to the omitted material from the "regression analysis" class and the "car discounts" example.

6. Modelling: Which Variables "Belong" in the (Current) Model?

How strong is the evidence that each explanatory variable has a non-zero coefficient (i.e., plays a predictive role) in the current model?

- Process: "Regress" the dependent variable onto the (current) set of explanatory variables. For each explanatory variable, examine the "significance level" (synonymously, the "p-value") of the sample data with respect to the null hypothesis that the true coefficient of that variable is zero.
  - The closer the significance level is to 0%, the stronger is the evidence against that null hypothesis (i.e., the stronger is the evidence that this variable has a true nonzero effect and does indeed belong in the current model).
- Example: How strong is the evidence that Mileage, Age, and Make each "belong" in the model which predicts Costs from all three?

Which Variables "Belong" in the (Current) Model?

- How strong is the evidence that Mileage, Age, and Make each "belong" in the model which predicts Costs from all three?
  - For Mileage and Age, "overwhelmingly strong".
  - For Make, a "little bit of supporting evidence", but not even "moderately strong".

Regression: Costs				
	constant	Mileage	Age	Make
coefficient	107.340945	29.6477024	73.9582688	47.4337242
std error of coef	82.0422871	3.91510733	17.9148891	28.9836595
t-ratio	1.3084	7.5726	4.1283	1.6366
significance	21.7429%	0.0011%	0.1677%	12.9983%
beta-weight		1.1531	0.5597	0.2193
standard error of regression		48.9578919		
coefficient of determination		84.90%		
adjusted coef of determination		80.78%		
number of observations		15		
residual degrees of freedom		11		
t-statistic for computing				
95%-confidence intervals		2.2010		

- With more data, if the true coefficient of Make is non-zero, the significance level will move towards 0%, and the evidence for inclusion will be stronger.
- With more data, if the true coefficient of Make is really zero, the significance level will stay well above 0%, and the estimate of the coefficient will move towards 0 (the "truth").
- … and we've now discussed everything but the t-ratios.

Don't Concern Yourself with the "t-ratio"

- The t-ratio of a variable is simply its estimated coefficient, divided by the standard error of the coefficient. It tells you how many standard deviations away from 0 the estimate of the coefficient is.
- In earlier days, folks used a simple rule of thumb:
  - "|t-ratio| > 2" means "strong evidence the variable belongs in the model" (really, means that the significance level is roughly 5% or less).
- Today, computers can compute the significance level directly, and the t-ratios are displayed only for historic purposes (i.e., to keep old statisticians happy).

Regression Analysis: How to *DO* It

Example: The "car discount" dataset

The slides marked with this symbol will be skipped during our first discussion of this dataset. After we cover "hypothesis testing," we'll return to them.

Hypothesis Testing

Regression: Discount				
	constant	Age	Income	Sex
coefficient	1971.72565	9.48991379	-0.035313	446.294355
std error of coef	146.147064	3.6320188	0.00366827	64.5567912
t-ratio	13.4914	2.6128	-9.6266	6.9132
significance	0.0000%	1.0423%	0.0000%	0.0000%
beta-weight		0.1746	-0.6735	0.4150
standard error of regression		301.19175		
coefficient of determination		69.68%		
adjusted coef of determination		68.74%		
number of observations		100		
residual degrees of freedom		96		
t-statistic for computing				
95%-confidence intervals		1.9850		

Note: 0.0000% means "positive, but very tiny."

The evidence for including Income and Sex in the model is overwhelmingly strong. The evidence for including Age is very strong.

Significance

- The significance level deals with the *marginal* contribution of a variable to the current model.
- Adding an irrelevant explanatory variable to a regression model will slightly increase the adjusted coefficient of determination about half the time. The significance level tells us if the coefficient of determination went up by *enough* to argue that the new variable is relevant.

Tests involving Coefficients

- In the full model, how strongly does the evidence support saying, "$\beta_{Sex} \geq \$200$"?
  - H_0: $\beta_{Sex} \leq \$200$, significance 0.01204% (overwhelmingly strong evidence against H_0, hence supporting original statement)

446.294	estimate/prediction of unknown quantity
64.557	measure of uncertainty
100	sample size
3	number of explanatory variables in regression, or 0 if dealing with a population mean

Null hypothesis:				significance level of data with respect to null hypothesis
true value	≥	200		100.00000%
	=			0.02408%
	≤			0.01204%

(from t-distribution with 96 degrees of freedom)

From Session-2's "Hypothesis_Testing_Tool.xls"

Tests involving Coefficients

- Other statements?

Statement	Significance level of data (with respect to opposite statement)	Strength of evidence supporting statement
$\beta_{Sex} \geq \$200$	0.01204%	overwhelming
$\beta_{Sex} \geq \$300$	1.28444%	very strong
$\beta_{Sex} \geq \$350$	6.95385%	somewhat strong
$\beta_{Sex} \geq \$400$	23.75235%	quite weak

The significance levels changed as the single "test" cell in the hypothesis testing tool was varied.