Stat 102

Introduction to Business Statistics
Class 3

Richard P. Waterman

Wharton

Table of contents I

Today's module

Topics to be covered in this module:

- Last time
- Confidence intervals
- Hypothesis tests
- P-values
- Summary
- Next time

Last time

Main points:

- The Empirical Rule
- Sample-to-sample variability, standard error
- Central Limit Theorem

Facts that would be very helpful for Team 2_24 to believe

If

1. Sample proportions come from a Normal distribution, centered around the hypothesized value
2. The standard deviation of this Normal Distribution is about 0.061.

then we can use the Empirical Rule to judge how unusual a proportion that is 0.951 standard deviations below 0.24 really is.

What the group was relying on for their inference

1. The sample to sample variability of the proportions (0.061) so they could calculate their z-score
2. The reference Normal distribution, against which they compared their z-score

This is a conundrum: we don't have access or the ability to check the ideas we were relying on, because we only have a single sample.

The standard error of the mean

Solving the conundrum: The fundamental insight of statistics, is that even with a **single** sample we are able to estimate sample-to-sample variability. If we denote the standard deviation as σ and the sample size n, then

$$SE(\bar{X}) = \frac{\sigma}{\sqrt{n}},$$

and we estimate the standard error as

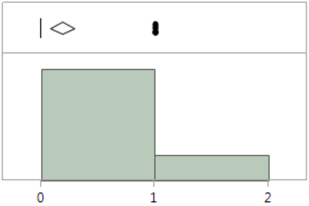
$$se(\bar{X}) = \frac{s}{\sqrt{n}}.$$

Where did the "s" come from?

- The raw data for team 2_24 is n = 55 with ten blues.
- In a spreadsheet enter this data, ten 1's and 45 zeroes (see *Group_2_24_raw_data.jmp*).
- Calculate the summary statistics:

Figure 1: Sample statistics for group 2_24

Is Blue?



Quantiles

100.0%	maximum	1
99.5%		1
97.5%		1
90.0%		1
75.0%	quartile	0
50.0%	median	0
25.0%	quartile	0
10.0%		0
2.5%		0
0.5%		0
0.0%	minimum	0

Summary Statistics

Mean	0.1818182
Std Dev	0.3892495
Std Err Mean	0.0524864
Upper 95% Mea	0.2870471
Lower 95% Mean	0.0765893
N	55

The Central Limit Theorem

Recall: the single group won't be able to look at the histogram of sample proportions to see whether or not they are Normally distributed (because they only have a single sample and you can't build a histogram from that one observation)

But, the **Central Limit Theorem** says that they don't have to.

The CLT states that with a sufficient sample size and an *i.i.d* sample, the sample mean must be **Normally** distributed:

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right),$$

where $\bar{X}$ is the sample mean, μ is the expected value of X and σ its standard deviation

The Central Limit Theorem, ctd.

In summary:

- It doesn't matter what the distribution of the random variable X is, so long as the sample size is large, $\bar{X}$ is always going to have a Normal distribution
- Proportions are special cases of means. They are just the mean of a set of $0/1$ data, so the CLT applies to the sample proportion too
- Sample proportions will be Normally distributed for sufficient sample size, n
- We want $n\,p$ **and** $n\,(1-p) \geq 10$ to believe that the Central **Limit** Theorem has *kicked in* for a Binomial random variable

Interval estimates

- What do you think the S&P is going to close at the end of this year?
- You could give a single number as an estimate, but you would almost certainly be wrong.
- You could give a range of numbers, which is more realistic.
- Or, you could give a range of numbers with a statement which conveys your confidence in the interval itself.
- The last one of these is the most informative and we will formalize it through the idea of a *confidence interval*.

Confidence Intervals

- What are they?
  - 1. A range of feasible values for an unknown population parameter, e.g. μ (population mean), p (population proportion) or ρ (population correlation).
  - 2. A statement conveying the **confidence** that the range of feasible values really does include the unknown population value.
- Where do they come from?
  - Inverting the Empirical rule.
  - If 95% of the time the sample mean is within $+/-$ 2 standard errors from μ, then 95% of the time the true and unknown μ is within $+/-$ 2 standard errors from the sample mean.

Usage

- Why are they important?
  - Move away from a single "estimate" to a range of values, which is more realistic.
  - Get to make the meta-level statement – our confidence about the first statement.
- How do I use one to make a decision?
  - Example from **DLOM.jmp**, is -16.5 a **feasible value** for the true mean of the DLOM percentage?
  - Answer: look to see if -0.165 lies in the confidence interval.
  - If it's in the interval then it's a feasible value.
  - If it's outside the interval then it is not feasible.

Three levels of information

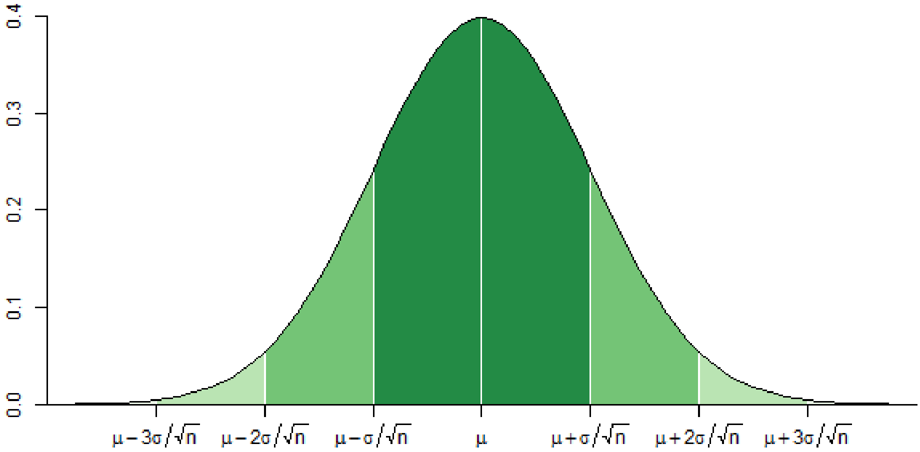
Level	Statement	Terminology
Entry level	I think that the average DLOM percentage is -16%.	Point estimate.
Better	I think that the average DLOM percentage is between -17% and -15%.	Interval estimate.
Best	I think that the average DLOM percentage is between -17% and -15% **&** I am 95% confident in this statement.	Confidence interval.

Confidence interval for the mean

Tell me where you think the true mean (μ) lies.
Before we have collected any data we know a lot about the sample mean (from the Central Limit Theorem).

- Sample means are approximately normally distributed.
  - $\mathrm{E}(\overline{X}) = \mu$.
  - $\mathrm{Var}(\overline{X}) = \sigma^2/n$.
  - $\mathrm{SD}(\overline{X}) = \mathrm{SE}(\overline{X}) = \sigma/\sqrt{n}$.

The sampling distribution of the sample mean



The key idea

- By the CLT and the Empirical Rule, we believe there is a 95% chance that a new $\overline{X}$ will be within $1.96\sigma/\sqrt{n}$ from μ:

$$P\left(\mu - 1.96\frac{\sigma}{\sqrt{n}} \leq \bar{X} \leq \mu + 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95.$$

- We don't know what μ is, but we can invert the previous statement to say that there is a 95% chance that μ is within $1.96\sigma/\sqrt{n}$ from $\overline{X}$.
- Putting this statement into a formula:

$$\overline{X} \pm 1.96\sigma/\sqrt{n}$$

  provides a 95% confidence interval for μ.
- This is the range of feasible values.

Confidence Intervals do not have to be 95%. 95% is just a convention. If you want something else (say 90%) then switch the [1]**1.96** to **1.645**.

[1]Don't forget that the exact value for the 0.975 normal quantile is 1.96, which we often round to 2 for convenience

The t-distribution

- There is a practical problem with the confidence interval formula $\overline{X} \pm 2\sigma/\sqrt{n}$.
- It contains σ which is almost always unknown.
- The natural thing to do is replace it with s, which we will do.
- This adds additional uncertainty into the calculation.
- It leads to the introduction of the Student's t-distribution, which is very similar to the normal distribution, but the t has fatter tails.
- As the sample size n gets larger the t-distribution converges to the standard normal distribution.
- So this detail only matters for small sample sizes.
- There is an argument that with small sample sizes, you should not be doing inferential statistics in the first place!

Summary of the t-distribution

- T looks like Z, but with fatter tails.
- The 97.5 percentile cut-off for the T is therefore greater than 1.96 (that for the Z).
- t-tables to find these cut-offs are in the back of the text book.
- For small degrees of freedom, e.g. 3, there can be a large difference between the cutoffs (3.18 v. 1.96).
- By the time the degrees of freedom reach 30, the difference between cutoffs is very small (2.04 v. 1.96).
- In most practical problems this difference is not important and simply rounding the cut-off to 2 is a reasonable rule of thumb.
- When software is used to do the calculations, exact cut-off values from the relevant t-distribution are used.

Example from Discount for Lack of Marketability (DLOM) data

An approximate 95% CI for the population mean μ is given by

$$\overline{X} \pm 2 \frac{s}{\sqrt{n}}.$$

- Define the variable of interest as `DLOM percent difference`.

Summary Statistics	
Mean	-0.156
Std Dev	0.172
Std Err Mean	0.004
Upper 95% Mean	-0.148
Lower 95% Mean	-0.163
N	2044.000

Using the confidence interval to make a decision

- Is there evidence to reject that the DLOM percent difference is equal to -0.165?
- The interval (-0.163, -0.148) does not include -0.165.
- Therefore there is evidence at the 95% level of confidence that the population mean is not equal to -0.165.

A confidence interval for a proportion

- A pharmaceutical company needs to compare the performance of its clinical trials to an industry benchmark.
- One way of measuring this is to look at the proportion of trials that move from Phase I to Phase II.
- The *attrition rate* measures the proportion of trails that **fail** to make it to Phase II.
- The industry bench mark is 45%.
- Provide a 95% CI for the attrition rate and comment on whether there appears to be a concern with this particular company's trials.

Key insight

If we represent the outcome of each trial which is either Terminate or Continue as

$$\begin{aligned} Continue &= 0 \\ Terminate &= 1 \end{aligned}$$

Then the sample mean of the $0/1$ variables is just the sample proportion of Terminated trials.
This implies that the sample proportion is just a special case of the sample mean, and we can apply the Central Limit Theorem to the sample proportion itself.
In summary, everything we have learned to date still applies here!

Notation

- Call p the population proportion.
- Call $\hat{p}$ the sample proportion.
- Under appropriate conditions the sampling distribution of $\hat{p}$ is approximately:

$$\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right).$$

- Key assumptions:
  1. We have an independent sample from the population.
  2. Sample size is large enough for the Central Limit Theorem to have *kicked in*.
  3. One rule for sample size: both $n\hat{p}$ **and** $n(1-\hat{p}) > 10$.
  4. If this sample size condition does not hold there are exact tests that can be used (they are not a part of this course).

An approximate 95% CI for the population proportion

- An approximate 95% CI for the population proportion is given by

$$\hat{p} \pm 2SE(\hat{p}) = \hat{p} \pm 2\sqrt{\frac{p(1-p)}{n}}.$$

- Unfortunately we do not know p so we replace it with $\hat{p}$ in the standard error calculation (just like replacing σ with s).
- So we use:

$$\hat{p} \pm 2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

The attrition data

- There were $n = 48$ trails.
- Of these, 30 were terminated, so $\hat{p} = 0.625$.
- Checking the sample size conditions $n\hat{p} = 30$ and $n(1 - \hat{p}) = 18$.
- The approximate 95% CI is given by

$$0.625 \pm 2\sqrt{\frac{0.625 \times 0.375}{48}} = 0.625 \pm 2 \times 0.070 = (0.485, 0.765).$$

- Reporting the interval on a percentage basis gives (48.5%, 76.5%).
- Notice that 45% is **not** in this interval.
- There is clear evidence that this company is not keeping up with the industry benchmark. Its attrition rate is significantly higher than the benchmark.

Reconciling the JMP output

The JMP confidence interval output is slightly different because

1. It used the t-distribution and not the normal to replace 2 with 2.0117.
2. It used (n - 1), 47 in the denominator of the standard error.
3. Neither of these are important and the discrepancy reduces as *n* gets larger.

Summary Statistics	
Mean	0.625
Std Dev	0.4892461
Std Err Mean	0.0706166
Upper 95% Mean	0.7670622
Lower 95% Mean	0.4829378
N	48

Interpreting a confidence interval

- What does the 95% confidence in the interval really mean?
- You must **not** say

  There is a 95% probability that μ lies in the interval.

- You can't say this because μ is not a random variable, just some fixed but unknown number. Therefore there are no probability statements to be made about it.
- The 95% is a property of the procedure, not a specific interval.
- You can say

  95% of intervals created according to this procedure are expected to contain μ.

- In practice you get a single interval and you act as if it contained μ.
- The word *confidence* used in this context is a *term of art*.

The goal of the hypothesis test

- A hypothesis test allows us to make a decision between **two** options.
- They are very similar to Confidence Intervals in their construction.
- Unlike a CI they provide a measure (p-value) of the weight of evidence against the null hypothesis.
- This p-value can provide more force to your conclusions.

The decision making framework

- Deciding between one of two choices.
- Null hypothesis (H_0): status quo.
- Alternative hypothesis (H_1): the converse of the null. Sometimes called the *research hypothesis*.
- Example; jury trial. Null is **Innocent**. Alternative is **Guilty**.
- Note – the Null is taken as true *a priori*.
- Decision based on collecting data – the jury votes. If jury votes $= 12$ then **convict** else **acquit** and declare **NOT GUILTY**. Note, do not declare innocent!
- We never accept a Null hypothesis. At most, we *fail to reject* it.

Type I and Type II errors

- Two types of error (recall false positives and false negatives in Quality Control/Diagnostic testing)
  - Innocent, but declare guilty (null true but go with alternative – Type I)
  - Guilty, but say innocent (alternative true but go with null – Type II)

	The Decision	
	Decision: innocence	Decision: guilt
Truth: Innocent (H_0)	✓	✗ Type I error (α)
Truth: Guilty (H_1)	✗ Type II error (β)	✓

The decision rule

- **Assuming** that H_0 is true then what is the probability that $\overline{X}$ is more than 2 standard errors from μ_0, the null hypothesis value?
- From the Empirical Rule this has a 5% chance of happening.
- Define the decision rule as:

  Reject H_0 if and only if $\overline{X}$ is more than 2 standard errors from μ_0.

- This rule has a 5% chance of incorrectly rejecting H_0 when H_0 is true. That is, it has a Type I error rate of 5%.

The significance level, α.

- If we use a cut-off value of 2 for the test then given the Null is true there is only a 5% chance of making a Type I error.
- We call this 5%, the *significance level* of the test and write the significance level as $\alpha = 0.05$.
- If you used a cut-off value of 1.645 then the significance level of the test would be $\alpha = 0.10$.

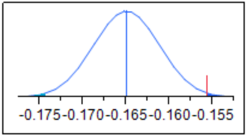
The one-sample t-test for the population mean

- Define the variable of interest as `DLOM Percent change`.
- Test: $H_0 : \mu_0 = -0.165 \qquad v. \qquad H_1 : \mu_0 \neq -0.165$.
- The test statistic for testing a single mean against a hypothesized value is called the *t-test statistic* and is calculated as
- $$\text{t-stat} = \frac{(\overline{X} - \mu_0)}{s/\sqrt{n}},$$
- where n is the sample size, s is the sample standard deviation, $\overline{X}$ is the sample mean and μ_0 is the hypothesized value under the Null.
- It is calculating how far away what we see $(\overline{X})$ is away from what we expect (μ_0), but on a standardized scale – the Empirical Rule scale.
- If what we see $(\overline{X})$ is a long way from what we expect (μ_0), then that is **surprising**.
- On observing a rare event, doubt the assumptions under which it is defined to be rare. That is, reject H_0.

JMP output for the one-sample t-test

Test Mean	
Hypothesized Value	-0.165
Actual Estimate	-0.1557
DF	2043
Std Dev	0.17207

	t Test
Test Statistic	2.4419
Prob > |t|	0.0147*
Prob > t	0.0073*
Prob < t	0.9927



The calculation for the t-test

$$\text{t-stat} = \frac{(\overline{X} - \mu_0)}{s/\sqrt{n}},$$

$$\text{t-stat} = \frac{(-0.1557 - -0.165)}{0.17207/\sqrt{2044}} = 2.4419.$$

This test-statistic is greater than $+2$ so we *reject* the Null hypothesis and conclude that the true mean is significantly different from -0.165.

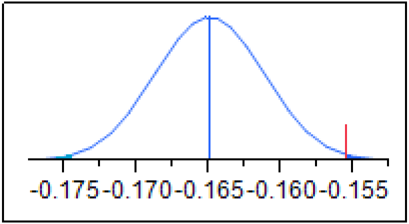
The p-value

- A measure of the credibility of the null hypothesis.
- Small p-values give evidence against the null.
- In English; the probability that if you did the experiment again and the null hypothesis were true, that you would observe a value of the **test statistic** as extreme as the one you saw the first time.
- It picks up the *repeatability* idea. If something is true (i.e. the null hypothesis) then you **should** be able to replicate the observed results. A small p-value says that it would be hard to replicate under the null hypothesis, hence the small p-value offers evidence against the null.
- An equivalent definition: the p-value is the smallest α-level at which H_0 can be rejected.

Graph to illustrate the p-value calculation

Mantra:

> *If the p-value is less than 0.05, then reject H_0 at the $\alpha = 0.05$ level of significance.*



The p-value is the sum of the (barely noticeable) dark blue shaded area to the right of the observed $\bar{X}$ (the red line) and its symmetric complement in the left tail.

Summary of the testing process

1. Set up the appropriate null and alternative hypotheses.
2. Identify the right test statistic.
3. Calculate the test statistic.
4. Compare the test-statistic to the [2]*cut-off value* or compare the p-value to α.
5. If the test statistic exceeds the cut-off value or the p-value is less than α then reject H_0.
6. Otherwise, fail to reject H_0.

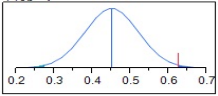
[2]The cut-off value comes from looking up the appropriate quantile in the t-tables, but we often round this value to 2 for simplicity when the test is two-sided and $\alpha = 0.05$.

- The industry benchmark for product attrition from Phase I to Phase II trials is 45%. How do we compare?
- Treat the data as 0/1 (1 = Attrition event) and note that the average **is** the proportion.
- Test: $H_0 : p_0 = 0.45 \qquad v. \qquad H_1 : p_0 \neq 0.45.$

Test Mean

Hypothesized Value	0.45
Actual Estimate	0.625
DF	47
Std Dev	0.48925

	t Test
Test Statistic	2.4782
Prob > |t|	0.0169*
Prob > t	0.0084*
Prob < t	0.9916



The p-value is 0.0169 and statistically significant, so there is indeed evidence for a difference from the benchmark.

The formula for the Z-test for a population proportion

From Stine 16.2:

$$Z = \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}}.$$

In the example:

$$\frac{0.625 - 0.45}{\sqrt{0.45 \times 0.55/48}} = 2.437.$$

- Why the slight difference between this and the JPM output?
- When we code as 0/1, JMP implements the t-statistic formula which is very similar to the Z-test above, but uses s in the denominator rather than $\sqrt{p_0(1 - p_0)/n}$.

Statistical significance and practical importance

- A highly statistically significant result does not imply a practically important result.
- A very small p-value allows you to strongly reject the null hypothesis.
- But that doesn't tell you if the alternative is a *meaningful* distance from the null.
- In medicine there is an identical concept of a *clinically meaningful* effect.
- But if you just think about practical importance without looking at statistical significance you run the danger of being a perpetual *noise chaser*.
- Good decisions consider both statistical significance and practical importance.

Module summary

1. Confidence intervals
2. Hypothesis tests
3. P-values

Next time

- Comparative analytics: comparing two groups