Stats E139 Fall 2015

Study Notes

David Wihl

October 14, 2015

Contents

1 Logistics

Prerequisites: AP Stats, Stats 101, 101, 102, 104 or 110.

Uses Calculus and includes algebraic derviations. Answers questions not covered by intro Stats class.

Office hours: SC-614, 7:30-8:30pm, also by appointment

Office (not used) 617-495-8711 or preferably via the stats dept 617-495-5496. Best of all use email: krader@fas.harvard.edu

TA: David Haswell

Combination of undergrad / grad level (master's level) material.

Sections: . Wed 5:30-6:30 Sever 112, followed by Office hours 6:30-7:30, Tory and Alina . Thu 7:40-8:40, 1 Story St 307, by David. Available on video, followed by Office hours

Lectures will **not** be broadcast live. There will be delayed video posted within 24 hours.

All material in lecture is tested even if it isn't in the lecture notes. The textbook is useful but not required.

R is used extensively. Course assumes no prior R knowledge. First homework requires some basic R. Start with "Comprehensive Intro to R" on slide 13.

Lectures and classroom is interactive - should bring questions.

Exams: final exam requires using R which is why it is a take-home exam.

Group project is 2-4 people, likely analyzing a data set. Final project grade: 2/3 group, 1/3 take home exam.

Homework is due by class time on Monday. Recommended to use LaTeX instead of Word. It is ok to scan or photograph hand-written work as long as it is submitted in PDF.

Collaboration is encouraged, but ensure that collaborators are cited appropriately.

There is no curve for exams and homeworks. At the end of the year, all students' grades are summed and a distrbution with A cutoff.

The mid-term on Nov 16 is through HW8, Intro to Linear Regression.

See Math Review sheet on Course Website.

Matrix Algebra will be emphasized later in the class to understand the math behind the techniques. Matrix Algebra will not be on tests.

For graduate students, there will be extra problems. Undergraduate students may do these additional problems for extra credit.

The Harvard College Stats139 class has three hours of lecture whereas this class has only two. So this class has the same bredth, but a lilttle less mathematical rigor and depth.

Statisticians are trained, whereas mathematicians are usually born.

A good statistician is able to communicate their analysis. This is often the most important step of the process.

Units 1, 2, 3, 4 are expected to refresher and will be done quickly.

Quantitative variables can be discrete or continuous.

Categories / Qualitative are not going to covered much in depth in this class.

1.1 R Demonstration

(slides 30 and 31)

Use weather data that is the same source, which is trustworthy. Use Max temperature per day because it is readily available unlike the average. Define what consistutes a "warmer" summer, by choosing an appropriate metric, like comparing means. Include a standard error. If the data is bell shaped, we can use a *t-test*. If the data is not bell shaped, we will explore other approaches.

(demo showing R Studio)

There is a lot more code in the Course site than the lecture notes because of the additional data processing and cleaning.

2 Unit 1 - Data Collection

(not many notes in this section as the lecture was primary a dramatic reading of the lecture notes).

Clinical experiments often has as few as 100 participants because the experiments are so expensive.

The balance of confounding factors by having random samples should isolate causation.

Key factors to an experiment: control group, randomized, and replicated. Usually 30 replications is sufficient.

People in a clinical trial may not be representative of the general population and followed regularly by a doctor. They may also be above average health which is why they choose to participate in trials.

One way to avoid confounding is to have *cross-over* groups, so they exchange methods during the course of the class.

Lecture stops at slide 19.

[a: rent or buy textbook
a: learn more R
a: evaluate R Studio]

Lecture 2 by Michael Parzen

Lecture restarts Unit 1, slide 19

Unless you do experimentation, you can't establish casuality. There are really only two sources of data: experimental and observational.

Types of Observational studies: (all time related) prospective, retrospective, longitudinal.

2.1 Sampling

The sample is a subset of the population of interest. Actual vs. Conceptual: you may not be able to get a respresentative sample, eg. Alzhemer's patients who are not institutionalized.

Parameters: μ population mean, π (or p) is true but unknown population proportion.

Statistics is a function of your data - a numerical summary of your sample.

Estimator $\bar{x}$, true μ.

Estimator $\overline{X}$, estimate is $\bar{x}$. Ideally a Simple Random Sample (SRS).

Census: sample everyone in the target population.

Sampling Frame: collection of units that are potential members of the sample. Bad sampling (**biased**): voluntary response, convenient sampling, question framing, confusing questions

2.2 Random Sampling

Systemic Random Sampling: every k member. Easy to admin, but population must be well mixed

Variable probability samplning: allow units to have unequal probabilities of being sampl

Interval validity: double check that allocation of groups is appropriate.

3 Unit 2 - Probability

3.1 General Probability Review

Probability is a measure of uncertainty and cannot be negative. (This section is mostly a recap so it is in summary form.)

Axiom 1: $P(A) \geq 0$

Axiom 2: $P(S) = 1$ guaranteed to happen

Axiom 3: If $A_1, A_2, ...$ are disjoint events, then

$$P\left(\cup_{j=1}^{\inf} A_j \right) = \sum_{j=1}^{\inf} P(A_j)$$

Sampling with replacement: n objects and making k choices, results in n^k possible outcomes.

Samplng without replacement: n objects and making k choices: $n(n-1)(n-2)...(n-k+1)$ possible outcomes.

A group of k people from a population of n people:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Conditional probability:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

$$P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$$

Bayes Rule:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Events are *independent* if $P(A|B) = P(A)$.

Parzen's 2x2 table:

	B	$\overline{B}$	
A	$P(A \cap B)$	$P(A \cap \overline{B})$	$P(A)$
$\overline{A}$	$P(\overline{A} \cap B)$	$P(\overline{A} \cap \overline{B})$	$P(\overline{A})$
	$P(B)$	$P(\overline{B})$	1.0

Summing in one dimension gives the marginals.

$$\begin{aligned} P(B) =& P(A \cap B) + P(\overline{A} \cap B) \\ =& P(B|A)P(A) + P(B|\overline{A})P(\overline{A}) \end{aligned}$$

Two events A and B are *conditionally indepedent given* E if and only if: $P(A \cap B|E) = P(A|E)P(B|E)$.

3.2 Random Variables

Two types: Continous and discrete random variables. Probability of a specific random variable = 0. Cumulative Distribution Function of a random variable given by $F_x(x) = P(X \leq x)$.

(See slides 15-18) Expectation, Properties of Expectation, Variance Definition, Properties of Variance.

3.3 Distrbutions

Normal / Gaussian Distribution (see slides)

In R, use the `pnorm(y, mean, sd)` command for CDF of the Normal.

Sums of Normal are Normal.

The Binomial is the most common discrete distribution.

See slide 26 for an approximation of the binomial distribution (not necessary in R).

3.4 Mean and Variance of $\bar{x}$

Because $\bar{x}$ is random, $E[\bar{x}] = \mu$.

Law of large numbers: $n \to \inf, var(\bar{x}) = 0, E(\bar{x}) \to \mu$.

Central Limit Theorm: with a large sample size, the distribution of $\bar{x}$ is normally distributed. Useful only if you are interested in $\bar{x}$, not the population.

Example 4: a) P(X > 70000). Can't be answered because we don't know the underlying distribution. b) $P(\bar{x} > 70000)$ can be answered.

4 Section 2 - Using R for Statistics and Homework Review

stopped section video at 18:00

Homework review notes: Internal validity - causal relational? Only valid when there is a randomized controlled experiment. External validity - will it generalize? Was the random sampling sufficient to generalize the conclusion?

5 Unit 3 - Hypothesis Testing

Section 1.3-1.6 in the text

If we see a relationship, we assume all confounding factors have been balanced so the malaria study, even with a very small sample, does have internal validity.

$$Y_{i,v} = \begin{cases} 1 & \text{if individual gets malaria when vaccinated} \\ 0 & \text{otherwise} \end{cases}$$

$$Y_{i,c} = \begin{cases} 1 & \text{if individual gets malaria when } \textbf{not} \text{ vaccinated} \\ 0 & \text{otherwise} \end{cases}$$

Treat as all independent, so Binomial distribution. Those variables should potentially follow the following distributions:

$$Y_{i,v} \sim \text{Bern}\,(P_v)$$
$$Y_{i,c} \sim \text{Bern}\,(P_c)$$

So what are the hypotheses?

$$H_0 : P_v = P_c$$
$$H_0 : P_v \neq P_c (P_v < P_c)$$

Presumably, malaria infections would not be increased by administering the vaccine.

Could use t-test only if distribution was normal or we had a large sample, neither of which applies.

Will use Fisher's Randomization Test.

Will need to determine proportion of infected to non-infected, ie $\hat{P}_c - \hat{P}_v$ or $(\overline{Y}_c - \overline{Y}_v)$, but it will be easier to measure $\overline{Y}$. If H_0 is true, the difference will be close to zero. If H_a is true, the difference will be large. But what will be sufficient difference to have statistical weight?

So divide by standard deviation to determine "how far is far" (like Z standardize). Since the dataset is so small, we will simulate more data.

5.1 The Hypothesis Testing Framework

Four steps in testing Hypotheses:

- Formulate hypotheses H_0 and H_A
- Calculate test statistic
- Calculate p-value based on a reference distribution of the test statistic (assume H_0 is true)
- Determine conclusion of the test by

A **scientific hypothesis** makes a testable statement about the observable universe.

A **statistical hypothesis** is more restricted. It concerns the behavior of a measurable (or observable) random variable. It is often a statement or claim about a parameter of a population or distribution.

These two types of statistical hypotheses for any scientific:

Null hypothesis (H_0) assumption specifying a possible truth, typically the absence of an effect. Presumes no change, old beliefs. Any discrepancy between the observed data and the hypothesis is due only to chance variation (noise).

Alternative hypothesis (H_a) assumption describing an alternative truth, typically some effect or some difference. A statement of possible new beliefs. An observed discrepancy between the observed data and the null hypothese is **not** due to chance variation.

5.1.1 Determine a Test Statistic

Statistic is a function of the data that summarizes it. $\hat{\mu} = \hat{\mu}(\mathbf{y})$

Test statistic a specific statistic used to weight evidence supporting or contradicting the null hypothesis.

Reference Distribution Probability distribution of the test statistic, assuming the null hypothesis is true $f(\hat{\mu}(Y)|H_0$. This is often called the *sampling distribution.*

Fisher's Exact Distribution is a way of doing this.

5.1.2 Calculate the p value

p **value** the probability of observing our test statistic or a more extreme one, assuming the null hypothesis to be true. Measure of strength of evidence of the hypotheses.

Calculated by *comparing* the observed test statistic to the reference distribution.

p **value is NOT the probability that the null hypothesis is true.** It is probablity of seeing our data on the null hypothesis. Use Bayes Rule to flip this around.

5.1.3 Significance Level of a Test

p **value** probability that the test statistic would be at least as extreme as *observed* under the null hypothesis

significance level (α) is the criterion compared against the p value. The null hypothesis is **rejected** if p value is lower than α. (Usually 0.05)

Generally, α reflects the probability of rejecting the null hypothesis when it is true (a Type I error)

Two sided are considered more conservative because it is usually considered harder to have a conclusive test.

5.1.4 Determining the Conclusion

We come to a conclusion about our hypotheses by comparing the $p$value to the Type I error rate.

If the p value $\leq \alpha$ we "reject the null hypothesis" and say the result is "statistically significant at level α"

If the p value $> \alpha$, we are "unable to reject the null hypothesis". This is different than concluding that the null hypothesis is true.

Based on the study design, we can then generalize internally and/or externally, which the **scope** of the inference procedure.

5.2 Fisher's Randomization Test

Randomization Test is for experiments - for casual inference.

Permutation Tests are for observational studies - for generalization.

In randomized experiments, uncertainty comes from randomness of an assignment. A Fisher Randomization Test is a *distribution free* test for treatment effect in **randomized experiments**. No assumption as to the underlying distribution has to be made. Generally, is $\mu_1 = \mu_2$?

The only assumption is Additive Treatment Effect (δ):

$$Y_{c,i} = Y_{v,i} + \delta$$

In practice, only the control or the experimental can be measured as one study unit cannot be both vaccinated and not vaccinated for example.

$H_0 : \delta = 0$, ie zero treatment effect for all units. Each unit's outcome is the same regardless of the treatment assigned. So the distribution would be identical in both groups.

$H_a : \delta \neq 0$ non-zero treatment effect for ALL units. (The textbook uses $Y* = Y + \delta$).

Assumptions:

- Random assignment to groups

- Under H_0, *independence* of study units. More precisely, there is *interchangability* of study units. The Y will be the same for a given study unit irrespective of treatment. This is taken advantage of when building the sampling distribution (through simulation).

5.2.1 Test Statistic

Difference of outcomes of the two groups:

$$\hat{\delta} = \overline{Y}_c - \overline{Y}_v$$

other test statistics could have been used, such as the difference between the medians.

Randomization Distribution is the reference distribution of a test statistic in a randomization test, where variation is due to random assignment of the treatment.

Procedure: Y is values of malaria. X is if someone is vaccinated. Simulate a new X^* with the same values as X but *in a different order*, leaving the same Y. Repeat multiple times to simulate and build the histogram. This will show if the distribution is extreme.

The number of simulations could be randomly very large. The maximum number of possible simulations is the binomial distribution:

$$\binom{t}{s}$$

where t is the total number of study units and s is the number where the effect was shown.

5.2.2 Calculating p value

One sided alternative:

$$\begin{aligned} H_0 :&\delta = 0 \\ H_a :&\delta > 0(\text{ or } \delta < 0) \end{aligned}$$

Be sure to justify the sides.

The p value is proportion of values above or below the observed test statistic.

$$\begin{aligned} \hat{\delta} =&0.5 \\ p =&P(\overline{Y}_c - \overline{Y}_v \geq \bar{y}_c - \bar{y}_v) \\ =&P(\overline{Y}_c - \overline{Y}_v \geq 0.5) = 0.029 \end{aligned}$$

so there is sufficient evidence to reject the null.

5.2.3 Conclusion

Under the significance level $\alpha = 0.05$ there is evidence that the vaccine is effective. More generally, if effect is NOT homogeneous, then there is evidence that the vaccine was effective for at least one volunteer.

So $\hat{\delta}$ means that the vaccine could be effective for 50% of the population.

Scope of inference: Internal Validity maybe not if the same hospital. External validity: possibly not, as these were volunteers.

5.2.4 Permutation Test

The Permutation Test is precisely the same although for observation studies.

This example of malaria was used only $\{1, 0\}$ as Y. In practice, Y will usually be a non-discrete number. However, X will always to be reduced to $\{1, 0\}$ to determine if study unit received the treatment or not.

See slides.

6 Intro to R

6.1 For loops

Use for repeated sampling, operating on a vector or matirx, Markov chains, simulation studies.

R has a `for` and `while loops`.

Example:

```
n.iter = 100
# create a vector of 100 NAs (like [float('nan')] * 100 in Python)
variable = rep(NA, n.iter)

for (i in 1:n.iter){
  samp = runif(10)
}
```

7 Unit 4 -T Based Inference

Textbook Chapter 2

7.1 The χ^2 distribution

Generally right skewed, values from 0 to ∞.

Let $Y = Z_1^2 + \ldots + Z_k^2$ where $Z_1, Z_2, \ldots, Z_k$ are i.i.d $\sim N(0,1)$ (standard normal distribution), then Y follows a χ^2 distribution with k degrees of freedom ($\sim \chi^2_{df=k}$). In English, a sum of standard normal variables squared, written as $Y \sim \chi^2_k$. This is related to the sample variance (since this is the square of standard normal distributed elements):

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{x})^2 \sim \frac{\sigma^2}{n-1}\chi^2_{df=n-1}$$

Need the σ in order to standardize.

The PDF is:

$$f(y) = \frac{1}{\Gamma(k/2)}(y/2)^k(1/2)e^{-y/2}$$

It supports $y > 0$.

Skew / shape: heavily right skewed.

The mean:

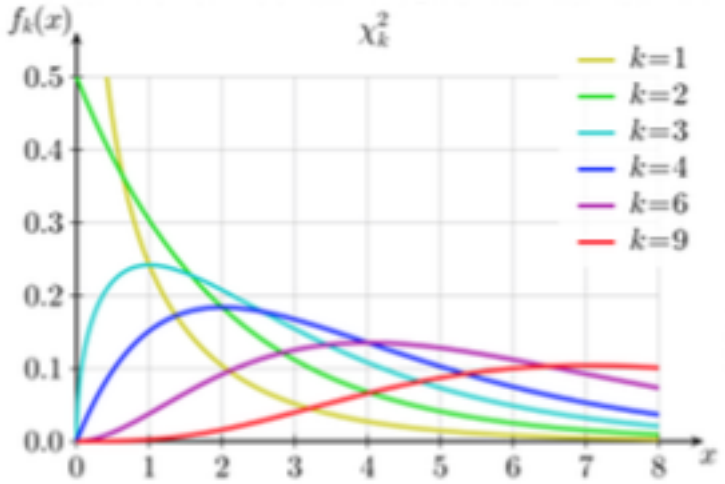
$$Y \sim \chi^2_{df=k}$$

$$\begin{aligned} E(Y) &= \int_0^\infty y \cdot f(y)dy \\ &= E(Z_1^2 + \ldots + Z_k^2) \\ &= E(Z_1^2) + \ldots + E(Z_k^2) \\ &= k \cdot E(Z_i^2) \\ &= k \end{aligned}$$

because each Z is normally distributed, the following applies:

$$\begin{aligned} Var(X) =& E(X - E(X))^2] \\ =& E(X^2) - [E(X)]^2 \\ =& E(X^2) - 0^2 = 1 \\ E(X)^2 =& 1 \end{aligned}$$

The expected value of a χ^2 distribution is the degrees of freedom.



As $k \to \infty$, the χ^2 distribution looks more like a normal distribution.

We care about χ^2 distribution because the distribution of the sample variance as a random variable is based on a χ^2 random variable when underlying observations are Normal.

For i.i.d, $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$, the sample variance is the r.v.

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2$$

then

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

We use $n-1$ instead of n because they are not completely independent. They are slightly negatively correlated, canceling out one degree of freedom.

To prove it, we need Gram-Schmidt orthonomalisation. (See Math23a!!! Also in Stat 111).

Use tables to lookup χ^2 values.

7.2 The t distribution

The t distribution has a slightly wider spread in the tails than Standard Normal. Also known as student-t distribution. This happens when you don't know the real σ and are forced to use S, which is less reliable.

$$\overline{X} \approx N(\mu, \frac{\sigma^2}{n})$$

allows us to Z-score

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

but we don't know σ so in practice we use

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

(which has two random variables ($\overline{X}$ and S). Now, take that previous definition of T and add $\sigma/\sqrt{n}$ to the top and bottom:

$$\begin{aligned}
T =& \frac{\overline{X} - \mu}{S/\sqrt{n}} \\
=& \frac{(\overline{X} - \mu)/(\sigma/\sqrt{n})}{(S/\sqrt{n})/(\sigma/\sqrt{n})} \\
=& \frac{Z}{\sqrt{S^2/\sigma^2}} \\
=& \frac{Z}{\sqrt{\frac{(n-1)S^2}{(n-1)\sigma^2}}} \\
=& \frac{Z}{\sqrt{\chi^2/df}}, \text{ since } \chi^2 = \frac{(n-1)S^2}{\sigma^2}
\end{aligned}$$

This generates a random variable with t distribution, which is standard normal, with k degrees of freedom. Since we are dividing by S instead of σ, there is greater uncertainty, which is why the distribution is slightly wider.

Recall $Z \sim N(0,1)$ and $\chi^2 \sim \chi^2_{df}$. It is a t-test because it is based on a t distribution. The t distribution critical values get closer and closer to the normal distribution (z) critical values as degrees of freedom increase.

This is important because there are a lot of times we don't know the true variance σ^2 but still want to know if the sample mean $\overline{X}$ matches the population mean.

7.3 One Sample t-based inference

One sample t-based inference follows the same hypothesis testing framework. We won't know the population mean or standard deviation (but assume it is approximately normal), but want to determine from the sample whether μ is particular value or not.

$$\begin{aligned} H_0 &: \mu = 70 (\mu_0) \\ H_a &: \mu \neq 70 \end{aligned}$$

$$t = \frac{\overline{X} - \mu_0}{s/\sqrt{n}}$$

Obviously if you have the real σ or μ, use that! But it will be a Z score, not a t test.

$$z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

Getting back to the t-test, with an example:

$$t = \frac{24923 - 24000}{22517/\sqrt{100}} = 0.410$$

$$\begin{aligned} p\text{-value} &= P(|t_{df=99}| > 0.410) \\ &= 2P(t_{df=99} > 0.41) = 2 * 0.3413 = 0.6826 \end{aligned}$$

0.3413 comes from R: `1 - pt(0.410, df=99)` (probabiity from a t distribution)

Since our p-value (0.68) is $> \alpha = 0.05$, we fail to reject the null. There is insufficient evidence to suggest financial aid has changed.

Note: t values are measures of *distance*. p values are measure of *probability*.

Even though the population may not be normally distributed, if there are sufficient samples, the samples will be normally distributed so the t distribution can be trusted.

Could also use R's `t.test()` to do this in one step, and also provides a confidence internval, which is an estimate for the true population mean μ.

7.4 The Confidence Interval

We want to make an *inference* about the population (e.g. μ or p) using information from the observed sample data. Instead of just a point estimate, it is more useful to have a *margin of error*.

CI is Estimate $\pm$ margin of error or Estimate $\pm$ (z value) $\times$ (SD of estimate).

We figure it out, using the formula from above:

$$t = \frac{\overline{X} - \mu_0}{s/\sqrt{n}}$$

and rearrange it to solve for μ_0:

$$\mu_0 = \overline{X} \pm t \cdot s/\sqrt{n}$$

(we use $\pm$ since we don't care if it is above or below $\overline{X}$).

At this point t is no longer a statistic - it is a critical value coming from the t distribution.

To find the 95% Confidence Internval for the **true mean** μ:

$$\bar{x} \pm t^*(s/\sqrt{n})$$

t^* has to be looked up in R. Use `qt()` (quantile coming from a t distribution) instead of `pt()`. `qnorm()` is the same as `qt()` with infinite degress of freedom.

If the confidence interval and α add up to one (e.g. 95% confidence interval $+\alpha(0.05) = 1.0$), then there is a direct relationship between CI and two-sided tests of hypotheses. When we reject the null, the calculated mean should be in the confidence interval. This may not hold for one-sided tests. (No one uses one-sided confidence intervals which is a value between x and infinity).

7.5 Caveats of t-based CI and Hypothesis Tests

Assumptions:

1. Observations are **independent**. This is true if the sample is randomly selected, but false if there is bias in the sampling.
2. **Normal Distributions** of the observations. This happens when the sample is large enough (via the Central Limit Theorm) or it is known that the population is normally distributed.

If these assumptions are not correct, none of the t-tests will work correctly.

7.6 Two Sample t-based Inference

This is the more common way to compare two different groups. Two different formulas for unpooled and pooled. Could also be paired.

Birthweight sample and smoking study. There was a lot of data, so there is likely correlations. The question is whether causality has been established. In this case, it was very specific, namely baby boys, survived at least 28 days and were single births. Good for confounding factors (internal consistency) but not good for generalization (external consistency).

In R:

```
cbind # combines columns
by # standard deviation, smoking
```

Anyone who did not answer the smoking question will be ignored, which is ok because there are only 10 out of 1200, so it isn't significant. If it was larger, we would hae to worry about non-response bias. If we cared, we could distribute them in the two groups. Assign them to smoking status's that makes them most and least and significant. If that does not materially change the results, then non-response bias has no effect in this study.

Let X_1 birthweight of a baby born from a smoking mother.

Let X_2 birthweight of a baby born from a non-smoking mother.

$$X_1 \sim N(\mu_1, \sigma_1^2) X_2 \sim N(\mu_2, \sigma_2^2)$$

$$\begin{aligned} H_0 :& \mu_1 = \mu_2 \\ H_a :& \mu_1 \neq mu_2 \end{aligned}$$

alternatively, to generate the statistic we want to use:

$$\begin{aligned} H_0 : \mu_1 - \mu_2 =& 0 \\ H_a : \mu_1 - \mu_2 \neq& 0 \\ \left(\frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \right) =& 0 \end{aligned}$$

however we don't the true population variances so we have to use S

$$\left(\frac{\overline{X}_1 - \overline{X}_2}{\sqrt{S_2^2/n_1 + S_2^2/n_2}}\right) = 0$$

becaues of properties of Expectations: $Var(X - Y) = Var(X) + Var(Y)$. We don't have to worry about covariances because these are supposed to be independent.

This will be an approximate t distribution because they are approximately normally distributed. The degrees of freedom will be based on both S_1 and S_2. To be conservative, we'll use the minimum of the two sample sizes $df = min(n_1, n_2) - 1$.

There isn't a causation here - there is only an association because it is an observational study. It also can't be generalized that well.

Confounding factors: mothers who choose to smoke might be unhealthy in other ways.

Confidence intervals work the same (see slide 33).

Unpooled test because we used two different variances. Pooled means you use a single variance.

R does this for you, however it has positive values. Make sure the order of variables match. When data is grouped, use: `t.test(Y ~ X, var.equal=T)`. If data is not grouped, use `t.test(Y1, Y2, var.equal=T)`.

7.7 Pooled Test

Assume a single variance:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2 | H_0)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

which is effectively combining the two sum of squares and then dividing by the combined variances. In this case, $df = n_1 + n_2 - 2$

Combining variance gives more statistical power but could make a significant mistake. Possible if the two variances are very close. In practice, everyone does the unpooled test. Technicality: the pooled test is truly normally distributed.

If the ratio of variances is > 2, use unpooled (always put larger variance in the numerator). If the ratio is < 2 you might be able to use the pooled test.

7.8 Paired t-test

There are pairs of observations for each subject in the study, e.g. take a reading on one day and another a week later. The two measurements should be correlated, in theory. By looking at the difference, we control for person-to-person variability. The column of differences is all we need in order to perform a one-sample t test to see if the differences are $= 0$.

In R, `t.test(x1,x2, paired=T)`. If the sample sizes in the two groups, you might be able to use the Paired t-test. If the sample sizes are different, you cannot use the Paired t test.

8 Section 3

Steps of Hypothesis testing:

1. State the null and alternative hypothesis
2. Choose a relevant test statistic
3. Calculate the observed test statistic, e.g. $t = |\bar{x}_0 - \bar{x}_1|$
4. Find the probability of observing the test statistic under the null (i.e. the p-value), e.g. $p(t \geq 0.5|H_0) = 0.03$
5. Conclusion: either fail to reject or reject, don't "accept." Mention scope of inference and internal validity. How general are the conclusions?

9 Section 4

One sample t-test, used when we want to test the mean of a single population. T stat has a df=n-1

$$t = \frac{\overline{X} - \mu_o}{s/\sqrt{n}}$$

$$t \sim T_{df=n-1}$$

test the mean of a population for a single parameter. "We think the mean of the population is this? Can we accept / reject?" The Student T is slightly more spread out than the standard distribution.

Paired t-test:

Same mechanics as 1-sample test (where single sample is the difference in the pairs). Used when there is some natural pairing between subjects in each sample (e.g. twins). $df = n-1$ (where each sample has n observations).

$$t = \frac{\bar{x}_{Diff} - \mu_{Diff}}{\frac{s}{\sqrt{n_{Diff}}}}$$

Use paired over pooled test when possible if you expect a correlation between the two samples.

Unpooled test:

Unpooled works irrespective how close the two variances. Unless you are really sure the variances are the same, don't use the pooled test.

$$df = min(n-1, m-1)$$

R would have a different df. Why is it a t distribution ?

$$\frac{Z}{\sqrt{\chi^2/n}} \sim T_n$$

Pooled Test:

only when you are really sure $\sigma_x^2 = \sigma_y^2, df = n + m - 2$. Larger df means more statistical power. But there is an additional assumption to worry about. Note different S_p^2 for pooled test and test statistic.

Assumptions:

in life, truly nothing is truly normally distributed. We are still making this assumption of normality. We are also assuming independence of observations. The Central Limit Theorm makes this more palatable. This is more about how the data was collected than any specific analysis. A good, randomized sampling method, means is more important than the math.

10 Unit 5 - Assumptions and Robustness of t-based Inference

Chapter 3 in the text.

10.1 Assumptions

We've seen four t-based assumptions so far:

1. One sample $t = \frac{\overline{X}-\mu}{s/\sqrt{n}}$

   If the assumptions are true, the sample will be normally distributed.

2. Two Sample (independent groups)

   - Unpooled $t = \frac{(\overline{X}_1-\overline{X}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_1^2}{n_2}}}$.

     If the assumptions are true, the samples will be *approximately* normally distributed. Also assumes that the two groups are independent.

   - Pooled $t = \frac{(\overline{X}_1-\overline{X}_2)-(\mu_1-\mu_2)}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$

     If the assumptions are true, the sample will be normally distributed.

3. Paired $t = \frac{\overline{D}-0}{S_d/\sqrt{n}}$

   If the assumptions are true, the sample will be normally distributed.

Recap of t-based assumptions:

1. Independence of Observations $X_i \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$
2. Observations come from a normal distribution
3. Two sample t-based Assumptions also assumes the two groups are independent. There is no covariance in either equation. We assume that the observations are independent both within the groups and with each other. $X_{1,i} \overset{i.i.d.}{\sim} N(\mu_1, \sigma_1^2)$ and independently $X_{2,i} \overset{i.i.d.}{\sim} N(\mu_2, \sigma_2^2)$
4. Pooled also assumes that the variances are the same. $\sigma_1 = \sigma_2$ and that the samples are independent.

10.2 Robustness

The performance of an inferential procedure when the assumptions fail is called **robustness**.

This is measured by:

Type I error probability of incorrectly rejecting the null hypothesis when it is actually true. False positive rate. Fixed typically $\alpha = 0.05$. If the assumptions are wrong, especially with small sample sizes, the probability of Type I errors increases dramatically.

Power probability of correctly rejecting the null hypothesis when the alternative is actually true. Usually, the alternative has a range of values so a value has to be picked. Defined as (1− Type II error).

Both types of error concern rejecting the null, but under different conditions.

10.3 Breakdown of Independence Assumption

The assumption requires independence within the group and between the groups. If between-group independence is violated, we can do a paired t-test.

Based on collecting data, if correlations are related within groups, then there is a breakdown of independence assumption.

For example collecting information about gender, but both samples of gender are from the same peer group. **Clustering** is when subgroups of units are similar to each other.

If they are centered around different averages, then regression will allow us to "control for" it.

There are extensions of linear regression to control for this grouping based on the type of correlation:

Serial effect Dependence over time

Spatial effect Interference across space

You can only take these clusters into account if you can measure how these clusters are close to each other either in time or space. Think carefully how was the data was collected. Did it all come from the same neighborhood?

When independence assumption is violated then the Sampling Variance calculations will be incorrect. When two results are always the same, the sample size

is artificially inflated that may lead to incorrect statistically significant results. More advanced calculations are needed or redefine units. (See textbook, chapter 15). Example: selecting NFL team stats over multiple years and treating as independent data, when it isn't.

Use of graphics and plots can help identify these clusters. Split the data into batches manually and see if there is correlation (e.g. boxplot, periodicity). Check across space and time to see if there are patterns.

10.4 Breakdown of Normality Assumption

Use visual examination of shapes (left or right skews, symmetry (long or short tailed), unimodal / multimodal) of the sample distribution. Does the sample distribution follow a normal distribution?

Overlay the density plot with the **kernel density** and the normal curve (for example, fitted by the method of moments, MOM). If the two curves fit reasonably well, it is likely a normal distribution.

In R,

```
x <- data$tuition[data$public==1]
hist(x, col="light gray", main="Public", xlab="Tuition, in $1000",
    prob=TRUE, breaks = 10, cex.main=2,cex.lab=2)
# Kernel density
lines(density(x, adjust = 2), col = "blue", lwd=2)
# Approx. normal curve, fitted using MOM
points(seq(min(x), max(x), length.out=500),
    dnorm(seq(min(x), max(x), length.out=500), mean(x), sd(x)),
    type="l", col="red", lwd=2, lty=2)
```

Use a boxplot show outliers.

Best of all, use a quantile-quantile (QQ) plot which should generate a straight line when both distributions are normal as well as chunking of points around 0 and less towards the extremities. Good for showing distributions away from normality that a histogram wouldn't necessarily show. When the sample size is small (e.g. 20), even a normal distribution will not look normal on a QQ plot. R: `qnorm()`, `qqline()`. QQ plots can contrast any distribution like χ^2 (see homework 3). Samples should follow a χ^2 distribution. Even when sampling from a normal distribution, the plot might be skewed, especially when the sample size is small.

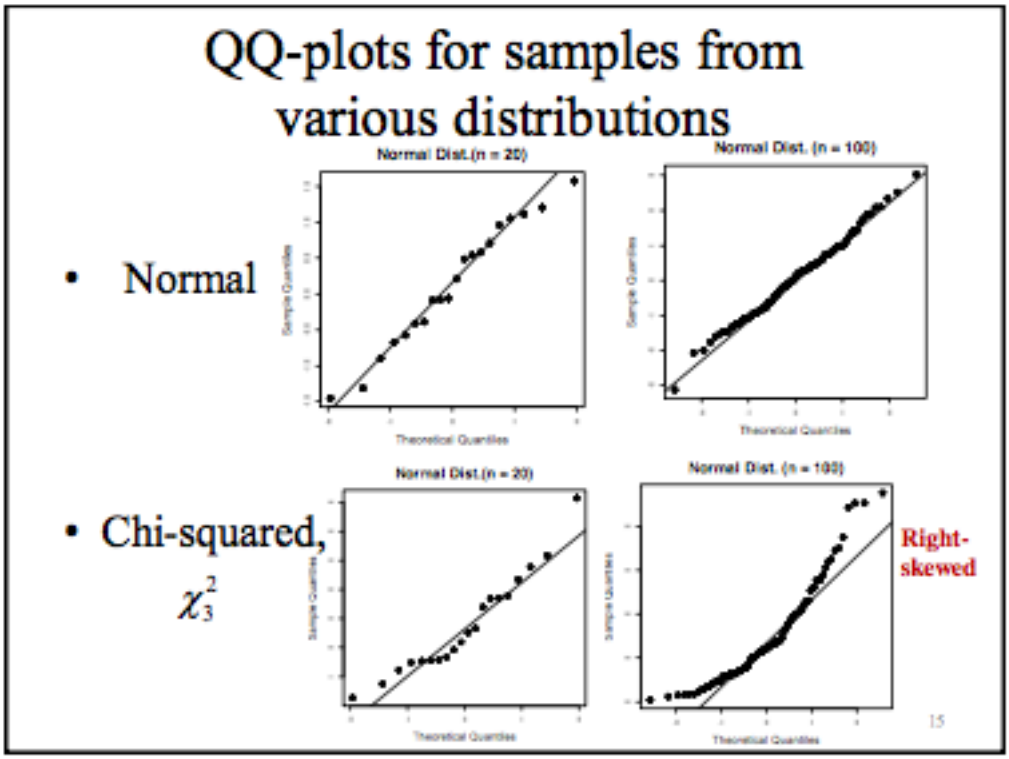


Figure 1: QQ Plots

10.4.1 Analytical Tests for Normality

There exists analytical tests for normality such as Anderson-Darling or Shapiro-Wilk. However, use **caution**! When there are only small sample sizes (e.g. < 20), these tests have low power and may not even reject for a non-normal sample.

For large samples, these tests will always reject, but we know from the Central Limit Theorm, that the larger the sample, the more likely it is to be normally distributed, irrespective of the population distribution.

The non-normality should be assess relative to the problem, which implies using plots.

10.4.2 Uniform Distribution

$X \sim Unif(a, b), a < b$ means that the values are evenly distributed in the interval $[a, b]$ so all values are equally probable.

Drawing the PDF, shows a rectangle with length $b-a$ and height $\frac{1}{b-a}$ (since all PDFs have to have an area $= 1$.) So $y = f(x) = 1/(b - a)$.

Standard Uniform Distribution: $a = 0, b = 1$.

10.4.3 The Universality of the Uniform Distribution

If you take random samples from a uniform distribution, the samples should be normally distributied $X_i \sim N(\mu, \sigma^2)$. Then if you calculate one sample t-test from that sample population $T = \frac{\overline{X}-\mu}{S/\sqrt{n}}$, the resulting T values should follow a t distribution $T \sim t_{df=n-1}$.

If you then generate p-values based on the t-test, the resulting p-values should have a **uniform distribution**. If the t-values do not the t distribution, then the resulting p-values will not be uniformly distributed, and our assumptions are incorrect.

See sample code for Unit 5 to see how to sample from a distribution to verify this.

Since $\alpha = 0.05$, we will reject 5% of the time. If 5% of the p-values are less than 0.05, then we can conclude that the test is valid. Changing sample sizes (to even small sample sizes) will not change the distribution of the pooled case.

10.4.4 Robustness of t-test to Departures from Normality

The two-sampled t-test is robust under moderate skewness as long as the sample size is large. (See Example 3 in Unit 5 code).

Sample goes for Exponential distribution. As long as the sample size is large (e.g. ≥ 50), the t-test will still be robust. (See Example 5, in Unit 5 code). When the sample size is small (e.g. 10), we get 6% of rejection, increasing our Type I errors. (See Example 6).

Note that mean and standard deviation had to be coerced into 0, 1 which isn't typical for other distributions like exponential.

In summary, t-tests are fairly robust to departures from normality, especially in large samples (CLT). When the sample sizes are not equal, t-tests are more sensitive to skewedness and long-tailedness. For small samples, t-tests are somewhat sensitive to markedly different skewedness in two groups. Watch out for outliers.

When normality assumption is violated:

- t-test is usually still valid, or
- Use data transformation, or
- Use non-parametric test (Unit 6).

10.5 Equal Variance Assumption

For the Pooled t-test, we need to verify the Equal Variance Assumption.

When sample sizes are equal, the pooled t-test is fairly robust to unequal variances.

When sample sizes are unequal, the pooled t-test is typically not valid for unequal variances; the unpooled t-test is a robust alternative.

When the equal varianance assumption is violated:

- Use unpooled t-test; or
- Use data transformation; or
- Maybe the populations are not comparable?

It isn't enough to compare the means of the two populations - it is more important to compare the lower tail of the population having equal distribution. E.g. minimizing the number of light bulbs that burn out.

(See Unit 5 code, example 7). Recall that the ratio of the variances should be < 2 (always put larger on top) in order to use the pooled test.

The larger group, with the larger variance give p-values closer to 1, whereas the smaller group with the smaller variance brings p-values closer to 0. See the formula for pooled t test:

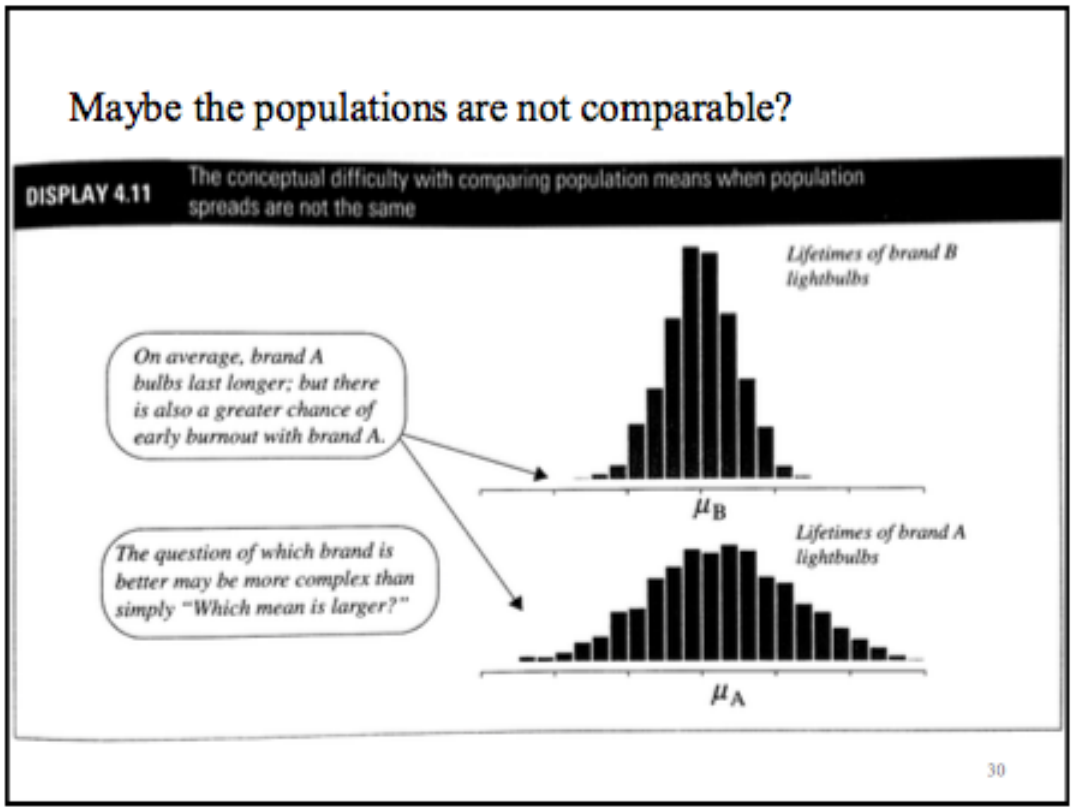


Figure 2: Maybe the populations are not comparible?

$$t = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Depending which sample we decide is X_1 or X_2 will affect the skewness of the p-value distribution.

Having a very high Type I error rate , like 37% (See Example 9) is very bad for the quality of our model. So, either make the group with the larger variance be X_1 or simply and preferably use an unpooled test.

10.5.1 Unpooled (Welch) Two-Sample t-test

This is what R uses to calculate degrees of freedom. Not needed for this class, but the formula is on Unit 5, slide 35.

10.5.2 Graphical Check of Equal Variances

Use box plots. Look out for outliers! (since variance is sensitive to outliers).

In practice, Statistiians use the unpooled test most often as it is more robust.

10.6 Formal F-test of Equal Variance

Start with graphical check and look at the *ratio of sample variances*.

Alternatively, use the F-test for comparison of two variances. In R, `var.test()`.

Same assumptions as for t-tests (i.i.d, Normality)

$$H_0 : \sigma_x^2 = \sigma_y^2$$
$$H_a : \sigma_x^2 \neq \sigma_y^2, \sigma_x^2/\sigma_y^2 \neq 1$$

Test Statistic:

$$F = \frac{S_x^2}{S_y^2} \sim F_{n_x-1, n_y-1}$$

Caution: will reject for large samples, even when the ratio is very close to 1.

10.6.1 F-distribution

Let $X \sim \chi^2_{n_x}$ and $Y \sim \chi^2_{n_y}$, independent of each other.

F is distributed as the ratio of the two variances.

$$F_{n_x-1,n_y-1} = \frac{\sigma_x^2 \cdot \frac{\chi^2_{n_x-1}}{n_{x-1}}}{\sigma_y^2 \cdot \frac{\chi^2_{n_y-1}}{n_{y-1}}}$$

Under the null hypothesis, $\sigma_x^2/\sigma_y^2 = 1$, so this simplifies to:

$$F_{n_x-1,n_y-1} \sim \frac{\frac{\chi^2_{n_x-1}}{n_{x-1}}}{\frac{\chi^2_{n_y-1}}{n_{y-1}}}$$

If this is true, then the following ratio has an F-distribution with n_x and n_y degrees of freedom.

$$R = \frac{X/n_x}{Y/n_y} \sim F_{n_x,n_y}$$

The above assumes a population. Generally, this is written as degrees of freedom for a sample, ie.

$$R = \frac{X/(n_x-1)}{Y/(n_y-1)} \sim F_{n_x-1,n_y-1}$$

As a ratio of variances, it will be positive, usually heavily right skewed.

(Later: very handy for ANOVA and model selection in Linear Regression)

(Don't worry about Two-sided p-values for F-test, Unit 5, slide 40).

11 Unit 6 - Transformations

Section 3.5 and Chapter 4 in the text.

11.1 Non-Linear (Log) Transformations

Linear transformations are in the form $ax + b$. Nonlinear transformations are square root, natural log, exponent, etc.

The log transformation is the most common transformation because it flattens exponential scales. It turns a multiplicative effect (which is very right skewed) and into an additive effect (which is more symmetrically skewed). It may also fix very large variance differences. This does not work for square root transformations.

Right skewed data cannot be used for a t test, whereas symmetric data can be subject to a t-test.

Most common choice is natural log transformation (for ease of derivation):

$$Z_i = ln(Y_i), i = 1 : n$$

Used when data is very skewed to the right ("positively skewed") or spread is larger in a group with a larger center.

The ideal result after the transformation is two symmetric samples with similar spreads but possibly different centers.

11.1.1 Interpretation - Randomized Experiments

Let T be the treatment group, and C be the control group.

$$\begin{aligned} Z_{T,i} &= ln(Y_{T,i}), i = 1 : n_T \\ Z_{C,i} &= ln(Y_{C,i}), i = 1 : n_C \\ Z_{T,i} &= Z_{C,i} + \delta, \text{ which is equivalent to} \\ \frac{Y_{T,i}}{Y_{C,i}} &= e^{\delta} \end{aligned}$$

For unit i responses, you have to exponentiate the difference:

$$exp(\overline{Z}_T - \overline{Z}_C) \text{ estimates } \frac{Y_{T,i}}{Y_{C,i}}$$

11.1.2 Interpretation - Observational Studies

$$Z_{1,i} = ln(Y_{1,i}), i = 1 : n_1$$
$$Z_{2,i} = ln(Y_{2,i}), i = 1 : n_2$$

For symmetric distributions, $E(Z_{j,i}) = \text{Median}(Z_{j,i})$. If the distributon is symmetric on the log transform scale, the means and medians are the same. When transforming back to the exponentiated scale, the mean will be messed up, but the medians will not be affected.

This is called *Monotoncity of logs*: $\text{Median}(\ln(Y_{j,i})) = \ln(\text{Median}(Y_{j,i}))$

Interpretation in terms of *a ratio of population medians*:

Let $m_1 = Median(Y_{1,j})$ and $m_2 = Median(Y_{2,j})$, then

$$exp(\overline{Z}_2 - \overline{Z}_1) \text{ estimates } \frac{m_2}{m_1}$$

So the median of the second population is $exp(\overline{Z}_2 - \overline{Z}_1)$ times as large as the median of the first population.

11.1.3 Log Transform for Paired t Test

$$Z_i = ln(Y_{2,i}) - ln(Y_{1,i}), i = 1 : n$$

take differences in logged outcomes within pairs **before** averaging!

In Randomized Experiments: $Y_{2,i}/Y_{1,i}$. $exp(\overline{Z})$ estimates a multiplicative effect treatment.

In Observational Studies: Let the median of ratios be $m(\neq$ ration of medians!), $\text{Median}(Y_{2,i}/Y_{1,i}) = m$, then $exp(\overline{Z})$ estimates m.

11.1.4 Graphical Interpretation

Look at distribution of log scale, QQplots. Outliers are not gone, but they are minimized.

11.1.5 Interpretation of Exponentiated Scale

Recapping the hypothesis test, we want to see the result on the exponentiated scale. Starting with log scale, we use a hypothesis test to check if the true (log) means are equal:

$$H_0 : \mu_1^* = \mu_2^*$$
$$H_a : \mu_1^* \neq \mu_2^*$$
$$t = \frac{(\overline{Z}_1 - \overline{Z}_2)}{\sqrt{\frac{S_1^{2*}}{n_1} + \frac{S_2^{2*}}{n_2}}}$$

($\overline{Z}$ is mean of data on log scale)

Example: The above hypothesis test produces a very low p-value and a (logrithmic) 95% confidence interval of (0.24, 0.72) for ratio of $\frac{m_1}{m_2}$. So the exponentiated confidence interval is $(e^{0.24}, e^{0.72})$ or (1.27, 2.05). This result makes sense because the interval does not 1. A multiplicative factor of 1 on the log scale is equivalent to a 0 in CI of non-log scale.

11.1.6 Limitations of the Log Transformation

Log scale does not work well for data with many small values [why?], especially zeros - best if all values are > 1.

$$Y_{j,i}^* = Y_{j,i} + \epsilon, Z_{j,i} = ln(Y_{j,i}^*)$$

If there are not many zeros, one can shift all observations by a small number (say, 0.01, or 0.1).

Cannot be used with negative data.

It doesn't matter what base of log is used - the shape of the transformation will always be the same. Sometimes log_2 or log_{10} are used to make interpretation better: allows for use of a doubling or a ten-fold increase for a one-unit change in the log-transformed variable.

11.1.7 Other Types of Transformations

If the log transformation is too strong (turns right skewed into left skewed), other transformations can be applied.

Square root or other polynomial

- Good for moderately right-skewed measurement data (e.g. counts, area sizes, etc.)

- Difficult to interpret results
- More appropriate for data values < 1 and > 1
- Using Y^2 if there is left-skewed data

Reciprocal taking $1/Y$

- For severely skewed data (waiting or failure times)
- Good for left skewed data
- Large values become small values.
- Often use negative reciprocal
- Can be used with negative data

Logit $log(\frac{Y}{1-Y})$, or $arcsin(2Y-1)$

- Good proportions or percentages to transform to real line
- Used with logistic regression

Could also flip axes to turn left skewed into right skewed data (e.g turn literacy rate into illiteracy rate).

11.2 Choosing a Transformation

Transformations, other than $log(Y)$ are difficult to interpret.

Choices are made by considering the nature of the data (e.g. counts, waiting times, proportions, etc.)

Plot the *transformed* data and assess where the result conforms with the assumptions of a chosen test (equal spread, normality, outliers, etc.).

Financial data often benefits from log transformations since the underlying data is often distributed in log-normal form (it is normal after taking the log).

11.3 Summary: Evaluating Validity of t-tools for a Problem

Evaluate independence assumption by considering the data collection method.

Use graphs to evaluate normality, similarity of shapes equality of variances and outliers

If needed

- transform data
- consider possible justifications for removing outliers
- or use robust t-tools (e.g. unpooled test)

Be cautious about removing outliers. Do not remove outliers unless there is justification to remove the outlier.

11.4 Alternatives to the t-tools

aka Nonparametric Tests based on Ranks.

Data is ordered and ranked from smallest to largest.

Use if sample sizes are small (e.g. < 30 for bell-shaped data or ever larger if skewed), or there are still outliers (and transformations did not help).

Transform all observations based on **ranks**. For two independent samples, use the **Rank-Sum Test** (aka Mann-Whitney or Wilcoxon Test).

For paired-samples, use the **Sign Test** or **Wilcoxon Signed-Rank test**.

11.5 Nonparametric Tests based on Ranks

11.5.1 Parametric vs. Nonparametric

Parametric procedure makes an assumption about underlying distribution of the observations (e.g. t-tests assume Normal Distribution).

Nonparametric procedure makes no assumption of the data generating process and therefore does not assume the observed data follows a specific distribution. Example: randomizaiton and permutation tests

11.5.2 Rank-Sum Test

Suppose we have two samples X and Y with different sizes drawn independently from two populations where $n_x \leq n_y$.

First the samples are transformed by **Rank** (e.g. sorted) into a combined set, Z. Ties (or duplicate values) are handled by averaging the corresponding Ranks (e.g. Rank 1, 2 becomes Rank 1.5)

Second, a permutation test is performed on ranks using the following test statistic:

$$T = \sum_{i_x=1}^{n_x} Z_{1,i}$$

which sums up the ranks for each group.

Then compare T for each group to see if they are close enough. If there is a difference in sample size, the rank sum has to be proportioned by sample size.

12 Section 6

HW5Q3, simulations. Frequentist approach.

```
# set up problem
n.sim = 10000

#initialize vectors with correct size
x1 = rep(NA, 50)
x2 = rep(NA, 50)
```

put in cheatsheet appropriateness of test - see section notes

Use R Markdown to embed R output into LaTeX