CEPC 0904
Brown University
Summer 2021
Professor Jack Hester

Probability/Stats Lecture Notes + Quiz Guide

Quiz 2 will focus on basic statistics and hypothesis testing (week 3 lecture videos). In these notes I will provide you with some information that will hopefully help you focus your studying and note content as well as clarify and add onto lecture content. You will be asked to determine which test is most appropriate given a study design. You will not need to write any R code for the quiz, but you might have to interpret the output of one of the tests listed below.

Probability Distributions

Probability distributions are important when determining what your data look like, what you might be able to expect, and what statistical tests are valid.

Bernoulli: Describes the outcome of a single trial (sample) where there are two possible outcomes that occur with probabilities p of "success" and 1-p of "failure" respectively.

Binomial: Describes the outcome of repeated Bernoulli trials (repeated sample where there are only two possible outcomes)

Normal: Describes a very common distribution of a continuous outcome and has very special and useful properties. It's more or less the continuous variable version of the Binomial distribution.

Poisson: Describes the probability of a certain number of events occurring over a certain interval of time or space.

Don't worry about exponential for the quiz.

Statistical Tests

The quiz will cover basics of t, z, and ANOVA tests. Below I will further define some terminology that I used in the videos, restate when you use certain tests, and state assumptions.

Terminology

Null hypothesis (H_0): A general "baseline" statement that there is no statistically significant result.

Alternative hypothesis (H_A): The statement or proposition that there is a statistically significant result (often difference between means).

Continuous variable: The variable can take on an any value (infinitely many values or "uncountable"). An example would be time, because you can have any value (any number of minutes, seconds, milliseconds, etc.). A person's weight would be another example.

Discrete variable: The variable can only take on a set ("countable") number of values. Examples include coin flip outcomes or money in your wallet (money can only take on a set number of values even though it might take you a while to figure out, because you can count only down to the cent). Note that, unlike categorical variables, discrete variables take on numeric values.

Categorical variable: The variable can be grouped (into "categories") that are defined by words rather than numbers. Examples include highest level of education, race or ethnicity, and gender. These variables can sometimes be arranged in some kind of order or ranking, which is referred to as "ordinal" variables.

Random sampling: The samples from the population are taken without any pre-defined pattern, decision, or logic (just like random numbers).

Outcome (dependent) variable: The variable you are predicting and for which you want to understand how different "upstream" factors impact it.

Predictor (independent) variable: A variable (sometimes one of many) you are using to predict what some outcome will be.

Standard deviation (σ): A way of measuring variability in a sample, closely related to variance (it's the square root of variance).

Independent populations: Two populations (from which samples are taken) are not related across any other variable. You can't ever know this absolutely, but you can use prior knowledge to make a reasonable assumption.

One sample t-test

The one sample t-test is used when you have one continuous outcome variable and one categorical predictor and you want to compare the mean of that sample to a known or reference mean.

H_0: There is no statistically significant difference between the true mean and comparison mean.
H_A: There is a statistically significant difference between the true mean and comparison mean.

Assumptions:

1. Continuous outcome variable
2. Outcome variable is normally distributed
3. Random sampling

Degrees of freedom:

The degrees of freedom = $n - 1$ where n is the sample size (number of observations).

Two sample t-test

The two sample t-test is used when you have two continuous outcome variables (from two samples) with the same categorical predictor (though the level doesn't have to be the same) and you want to compare the means between the two samples.

H_0: There is no statistically significant difference between the true means of the two populations.
H_A: There is a statistically significant difference between the true means of the two populations.

Assumptions:

1. Continuous outcome variable
2. Outcome variable is normally distributed (normality)
3. Random sampling (from both populations)
4. Populations are independent
5. *Variance of samples is equal (this does not need to be the case if you use a test for *unequal variances* such as Welch's)

Degrees of freedom:

For tests of equal variance and the same sample size, the degrees of freedom $= 2n - 2$ where n represents the number of samples from *one* population, i.e. if you have 20 individuals that were treated and 20 that weren't, then $n = 20$ and there are $40 - 2 = 38$ df. If you have unequal sample sizes or unequal variances, you will need to use a more complex equation (see `https://en.wikipedia.org/wiki/Student\%27s_t-test`). R and other statistical languages can perform these calculations for you as well.

Paired t-test

The paired sample t-test is used when you have two samples from the same population (such as a before and after) and want to compare means between the two samples.

H_0: There is no statistically significant difference between the true mean of the population at the different measurement points.
H_A: There is a statistically significant difference between the true mean of the population at the different measurement points.

Assumptions:

1. Continuous outcome variable
2. Differences between the matched sample pairs are normally distributed
3. Random sampling

Degrees of freedom:

The degrees of freedom $= n - 1$ where n is the number of pairs of observations, i.e. if you had 20 people before a treatment and the same 20 after, you would have 20 pairs and your sample size would be $20 - 1 = 19$ df.

One sample z-test

The one sample z-test is used when you have one continuous outcome variable and one categorical predictor and you want to compare the mean of that sample to a known or reference mean. You must also have more than 30 samples and know the true population standard deviation.

H_0: There is no statistically significant difference between the true mean and comparison mean
H_A: There is a statistically significant difference between the true mean and comparison mean

Assumptions:

1. Continuous outcome variable
2. Outcome variable is normally distributed
3. Random sampling
4. Population standard deviation is known

Degrees of freedom:

The z-test uses the z distribution, which does not depend on degrees of freedom, so you typically don't need to report df.

Two sample z-test

The two sample z-test is used when you have two continuous outcome variables (from two samples) with the same categorical predictor (though the level doesn't have to be the same) and you want to compare the means between the two samples. You must also have more than 30 samples and know the true population standard deviation for both populations.

H_0: There is no statistically significant difference between the true means of the two populations.
H_A: There is a statistically significant difference between the true means of the two populations.
Assumptions:

1. Continuous outcome variable
2. Outcome variable is normally distributed for both samples
3. Random sampling (from both populations)
4. Populations (and samples) are independent
5. Both populations' standard deviations are known

Degrees of freedom:

The z-test uses the z distribution, which does not depend on degrees of freedom, so you typically don't need to report df.

Paired z-test

The paired z-test is used when you have two samples from the same population (such as a before and after) and want to compare the means between the two samples, you have

more than 30 samples, and you know the true population standard deviation.

H_0: There is no statistically significant difference between the true mean of the population at the different measurement points
H_A: There is a statistically significant difference between the true mean of the population at the different measurement points

Assumptions:

1. Continuous outcome variable
2. Differences between the matched sample pairs are normally distributed
3. Random sampling (from both populations)
4. Both populations' standard deviations are known

Degrees of freedom:

The z-test uses the z distribution, which does not depend on degrees of freedom, so you typically don't need to report df.

One-way ANOVA test

The one-way ANOVA test is used when you have one categorical independent (predictor) variable with more than two levels and one continuous outcome variable. You can think of it as an extension of the two-sample t-test, but for more two categories of predictor variable.

H_0: There is no statistically significant difference between the true means of any of the populations/groups.
H_A: There is at least one population/group with a statistically significantly different mean than the others.

Assumptions:

1. Continuous outcome variable
2. Outcome variable is normally distributed for all samples
3. Random sampling (from both populations)
4. Populations (and samples) are independent
5. Variance across groups is the same

Degrees of freedom:

The overall degrees of freedom equals $n - 1$ where n is the number of observations. However, it can often be useful to know the number of degrees of freedom "between group" and "within group" or sometimes referred to as "treatment" and "error." You can read more about that here `https://people.richland.edu/james/lecture/m170/ch13-1wy.html`.

Two-way ANOVA test

A two-way ANOVA test is used when you have multiple categorical predictor variables (with different levels in each) and a continuous outcome variable, and you want to measure the impact of each categorical variable's levels on the outcome as well as the potential interaction between the two categorical variables.

H_{01}: There is no statistically significant difference in the means across levels of variable A.
H_{02}: There is no statistically significant difference in means across levels of variable B.
H_{03}: There is no statistically significant interaction between factors A and B.

H_{A1}: There is a statistically significant difference in the means across levels of variable A.
H_{A2}: There is a statistically significant difference in means across levels of variable B.
H_{A3}: There is statistically significant interaction between factors A and B.

Assumptions:

1. Continuous outcome variable
2. Outcome variable is normally distributed for all samples
3. Random sampling (from both populations)
4. Populations (and samples) are independent
5. Variance across groups is the same
6. *Equal sample sizes (not *always* required)

Degrees of freedom:

In a similar way to one-way ANOVA, the total degrees of freedom is $n - 1$, but the degrees of freedom between and among groups and for interaction effects that you might be interested in depend on the factors you're measuring across and how many levels they have. You can learn more about this process at `https://people.richland.edu/james/lecture/m170/ch13-2wy.html` or `https://www.youtube.com/watch?v=sHM8g4Y7G1s`.

p-Values and Confidence Intervals

p-Value: The probability that your result was due to random chance. Typically you compare this to your predetermined significance level. e.g., $\alpha = 0.05$.

Interpretation: We choose significant level $\alpha = 0.05$, and $p < 0.05$ then we say that there is a statistically significant finding so we rejet the null hypothesis. If $p > 0.05$ we say there is no statistically significant finding, and we fail to reject the null. (As you learned in "The ASA Statement against p-Values", this has some problems and isn't exact).

Confidence interval: A range of values that you expect to contain the true population mean, calculated as the point estimate (typically sample mean, $\bar{x}$) $\pm$ a margin of error (M.O.E.) which is determined by your significance level and distribution.

Interpretation: The confidence interval at $\alpha = 0.05$ means 95% $(1-\alpha)$ of the intervals we construct will contain the true population mean. Ex: "We are 95% confident that the interval (0.2, 2.3) contains the true population mean. If the C.I. contains (covers) 0, then we fail to reject the null hypothesis, but if it does not contain zero then we reject the null hypothesis.

NOTE: These metrics address the chance that we got our result due to random chance, but this does NOT mean that we are 100% certain of our result (or at least 95%). There is still the possibility that our groups differed at baseline for some other reason, whether we know what reason that is or not!