Lecture 3

Statistics Review

Kara McCormack

Thursday, January 18, 2024

Linear models are powerful!

- Linear models are incredibly powerful tool to model the relationships between variables.
- There are often times when linear models are not appropriate (e.g. for violating the technical conditions).
- A solid understanding of the linear model framework requires a strong foundation of the theory that goes into modeling.
- Today we'll review some of the ideas from introductory statistics.

How would you summarize a set of data?

- List the values, find a min/max, range, find the average
- Refer to a known distribution

Distribution (put loosely):

A mathematical rule that describes the relative frequency of different events (e.g., normal, uniform, Bernoulli, binomial, Poisson, etc.)

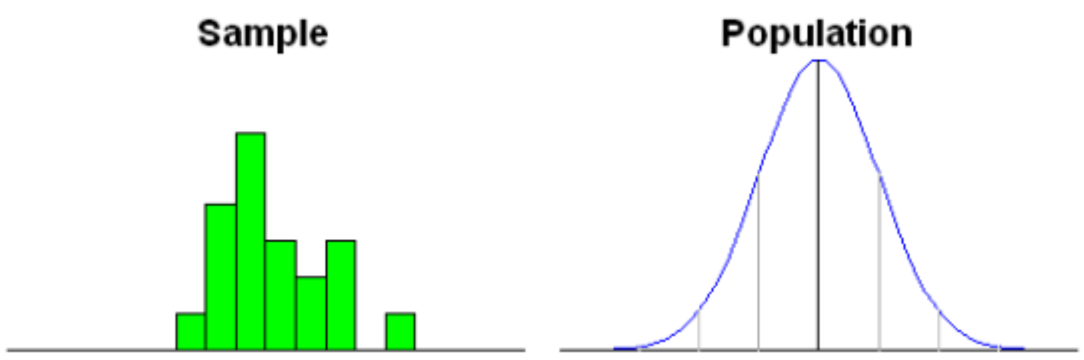
- For example, a normal distribution:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right]$$

For example

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right]$$

- We could calculate an average of all values: $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{N}$
- We could calculate a sample variance of the values: $s^2 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{(n-1)}$



How should we think about μ and $\bar{x}$?

- What is a variable?
- Recall from past math courses: $7 + x = 10$

Variable definition (loosely):

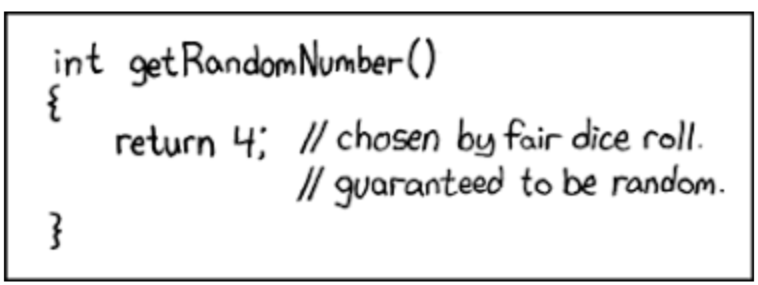
- A variable is a characteristic that can be measured and that can assume different values.
- For example: height, age, country of birth, grades at school

Random variable

- What is a random variable?

Random variables vs. realized values (loosely):

- A **random variable** is a number that moves around according to a probability distribution. In other words, a random variable is a way of mapping a random process to numbers.
  - Example: X = {1 if heads, 0 if tails} after flipping a coin.
  - Example: Y = sum of upward face after rolling 7 dice
- A **realized value** of a random variable is the value that is actually observed. (It's just a number, although it may have come from a probabilistic process.)

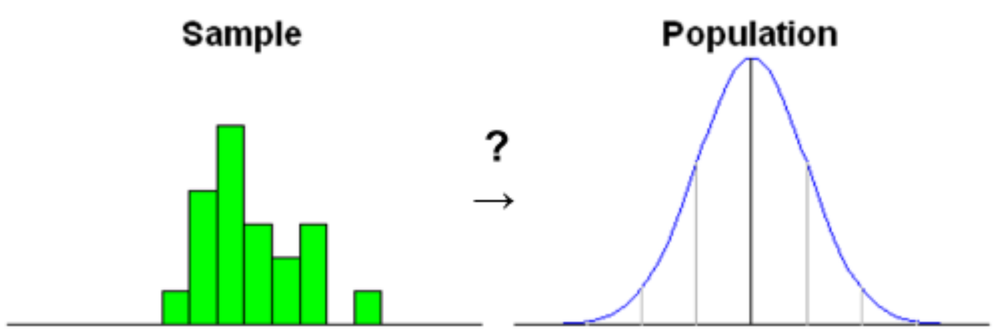


Distributions can refer to different sets of entities

- Population (e.g. all of humanity)
- Sample (e.g. the people in your study)

The Big Question

- Why is it possible to use what we’ve discovered about a sample distribution to say something about the population distribution? (And how would we go about this?)



The setup

- The set up: the population is too big to observe.
- We'd like to know the value of specific parameters, say for instance the mean of the population.
- We can't calculate the parameters directly. Instead, we observe a sample at random from a population.
- Based on the sample, we estimate parameters.

Samp*ling* distributions

Sample statistics are random variables.

- E.g., “the” sample mean is a random variable, but *your* sample mean is a realized value drawn from that random variable.

The behavior of random variables can be described by a distribution.

Distributions of sample statistics are called **sampling distributions**.

- Let’s look at an example!

Example of a Sampling Distribution

- Suppose we recorded the height of everyone in this class. Let's say the average height is 5' 8".
- Survey 10 classes at UNC, might get averages of 5'9", 5'8", 5'10", 5'9", 5'7", 5'9", 5'9", 5'10", 5'7", and 5'9.
- If you graphed all those averages in a histogram, you might get something like this:

Example (continued)

- If we surveyed all classes at UNC, (took the heights of each student in the class, recorded the mean in each class, and then made a histogram of those means) the distribution of the means would look more like a standard normal curve.
- Why does this happen? Answer: The Central Limit Theorem!
- Next, we'll talk about the Central Limit Theorem and the Law of Large Numbers.

Some (very light) Theory

The (strong) Law of Large numbers (loosely):

- As N goes to infinity, the sample mean converges to the population mean (or true value).
- Interactive example of Law of Large Numbers

The Central Limit Theorem (loosely): If our data are an unbiased sample, where each datum is independent and drawn from the same population, which has mean = μ and variance = σ^2, then, as N becomes large, the distribution of the sample mean will be:

1. Approximately Normal
2. centered on the population mean / true value
3. with variance: $\sigma^2_{\bar{x}} = \frac{\sigma^2}{N}$ (and thus SE: $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{N}}$)

Why would anyone take samples identically from the same population over and over again?

ANOVA!

- Analysis of Variance
- Used to compare the means of multiple groups.
- Example of one-way ANOVA: As a crop researcher, you want to test the effect of 3 different fertilizer mixtures on crop yield.
- The null hypothesis in a one-way ANOVA is that you sampled the same population repeatedly (i.e. the conditions don't differ)
- ANOVA uses the F-test to determine whether the variability between group means is larger than the variability of the observations within the groups. (More on this later.)

Repeated samples from the same population are useful!

- E.g. if we had a sample of data, and for a given population, knew:
  1. its mean,
  2. its SD, and
  3. that it was normally distributed,

then we could see how our sample compared to that population.

- We could do so by comparing our sample's mean to the distribution of sample means that would be drawn from that population under repeated sampling.

$z_{\bar{x}} = \frac{\bar{x}-\mu}{\sigma_{\bar{x}}}$ or $z_{\bar{x}} = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{N}}}$

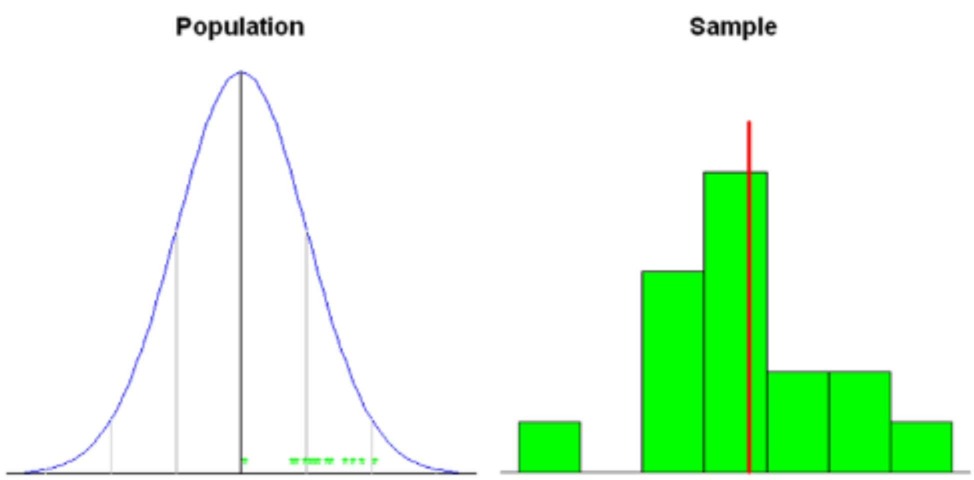
IQ example

Say we had the following IQ scores from students in a STATS 102 class:

$$IQ = \{118, 121, 101, 120, 113, 131, 126, 112,$$
$$116, 117, 124, 115, 120, 115, 120, 128\}$$

We know that the distribution of IQ scores:

1. has a known mean (100)
2. has a known SD (15)
3. is normally distributed

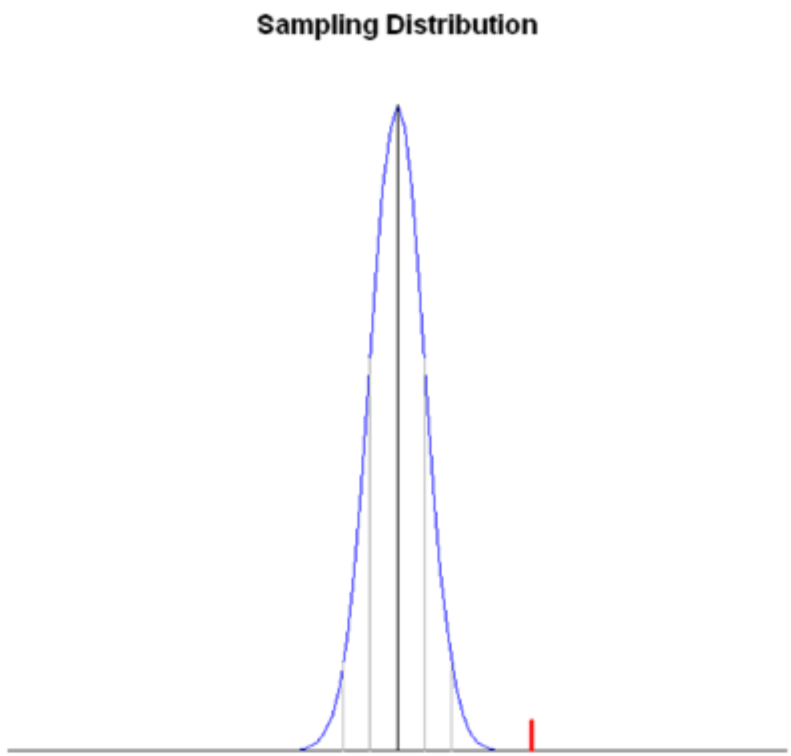


IQ example (continued)

$IQ = \{118, 121, 101, 120, 113, 131, 126, 112,$
$116, 117, 124, 115, 120, 115, 120, 128\}$

- Number of observations in sample: $N = 16$
- Sample mean: $\bar{x} = 118.6$
- Sample standard deviation $s_x = 7.1$
  - Recall $s_x = \sqrt{\frac{1}{(N-1)} \sum_{i=1}^{N} (x_i - \bar{x})^2}$
- Based on our data, we can calculate our Z score = $z_{\bar{x}} = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{N}}}$
- $19.84 = \frac{118.6 - 100}{15/4}$
- How does our value compare with a standard normal distribution?

IQ example (continued)



- 19.84!! That's pretty extreme.
- In fact, we can ask the question:
  - What's the probability of gathering a sample with a mean that far or further from the population mean? (The answer is the p-value!)

Null and alternative hypotheses

But first:

- **Null Hypothesis**: Denoted H_0, the null hypothesis is usually set up to be what is believed unless evidence is presented otherwise.
- **Alternative Hypothesis**: Denoted H_A, the alternative hypothesis is usually what we wish to show is true. It is more general than H_0, usually of the form the parameter is somehow not equal to the value used in H_0, without specifying exactly what we think it is.

But what really is an Alternative Hypothesis?

- Consider the brief video from the movie Slacker, an early movie by Richard Linklater (director of Boyhood, School of Rock, Before Sunrise, etc.)
- https://www.youtube.com/watch?v=b-U_I1DCGEY
- Watch from 2:22-4:30.

Reflecting on the video

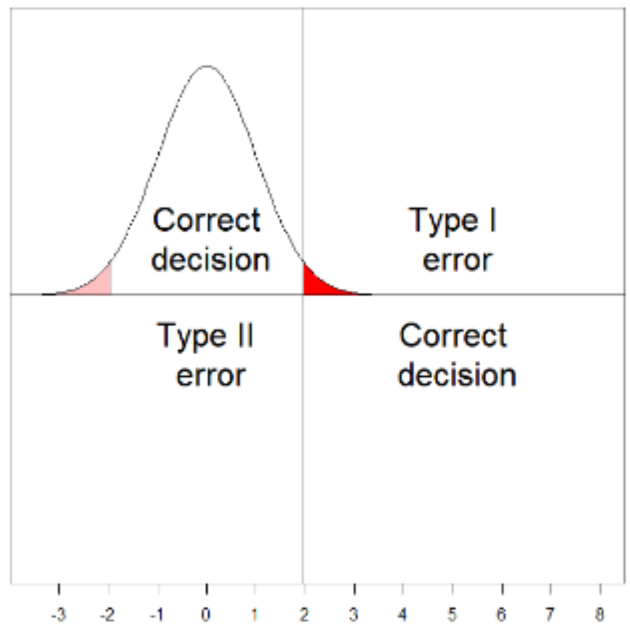
- The rider in the back of a taxi (played by Linklater himself) muses about alternate realities that could have happened as he arrived in Austin on the bus.
- What if he had found a ride instead?
- Could have taken a different road into a different alternative reality (and that in reality his current reality would be an alternate reality). And so on.
- What is the point? What is the relationship to sampling distributions?
- Since any procedure will have the potential to be wrong, we search for one that makes bad errors infrequently.
- There are two types of errors that can be made.

Two types of errors

- **Type I Error**: Rejecting H_0 when H_0 is actually true. Usually considered the worst possible error, and thus we find a procedure which makes the type I error with only a small probability (we denote this by α).
- **Type II Error**: Not rejecting H_0 when H_A is actually true. Having a small type II error is a secondary concern (after controlling the type I error). The probability of a type II error is denoted by β.
- $1 - \beta$ is known as the $power$.
- Let's check out a visual of this on the next slide.

Visualizing Type I and II Errors

You: In reality:	Fail to reject	Reject
Null TRUE	Correct decision	Type I error
Null FALSE	Type II error	Correct decision



P-value

- A p-value is the probability, if H_0 were true, of observing data as or more contradictory to H_0 if we were to repeat the experiment again.
- So if the p-value is .01, it means the data showed something that happens only about 1 time in 100 when H_0 is true.
- Considering that the particular set of data *was* observed, the reasonable conclusion is that H_0 must not be true.
- The rule is: reject H_0 if p-value < α.
- Th resulting test will have a type I error probability of α, which is the value we get to specify.
- α is often set to be 0.5.

The logic of hypothesis testing

There is an assymetry in the logic of hypothesis testing:

- Rejecting the null hypothesis conveys information.
- Failing to reject the null conveys *ambiguity*.
- A non-significant p-value does NOT allow you to accept the null.
- In other words, we cannot say: “There was no effect.”
- Instead, we say: “There was insufficient evidence to reject the null hypothesis.

What if you don't know the population SD?

- This is where the t-distribution comes in.
- Previously, we calculated our z-statistic because we knew σ.

$$z_{\bar{x}} = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{N}}}$$

- If we don't know σ, we can substitute it with the sample standard deviation, s:

$$t_{\bar{x}} = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{N}}}$$

- So, you use the sample SD as an estimate of the population SD, and compare the resulting test statistic to the t distribution instead of the normal Z distribution.

The t distribution

- The t distribution has only 1 parameter, its degrees of freedom (df).
- For a one-sample t-test, the df is $N - 1$ (i.e. the number of observations minus 1).
- Note that when the degrees of freedom becomes large $(> 30, > 50, > 100)$, the t-distribution becomes virtually indistinguishable from the normal.

$$f(t) = \frac{\Gamma(\frac{df+1}{2})}{\sqrt{(df)\cdot\pi}\Gamma\left(\frac{df}{2}\right)}\left(1 + \frac{t^2}{df}\right)^{-(df+1)/2}$$

What if we have 2 conditions?

- We can make a single statistic by subtracting the two means and comparing the difference to the sampling distribution of the difference between the means.

$$t = \frac{(\bar{x}_2 - \bar{x}_1) - (\mu_2 - \mu_1)}{s_{x_1 x_2}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$s_{x_1 x_2} = \sqrt{\frac{(n_1 - 1)s_{x_1}^2 + (n_2 - 1)s_{x_2}^2}{n_1 + n_2 - 2}}$$

where n_1 and n_2 are the sizes of each sample, respectively.

Parameter Estimation

We can also go in the other direction. That is, what population did our data come from?

- When drawing a sample from a given population, the most likely sample mean to get is the population mean.
- So, the sample mean is the “best guess” (point estimate) for the population mean.
- You can use the samp*ling* distribution to get a 95% confidence interval of the population mean.

Confidence Intervals

- If we know the population SD, then we can create confidence intervals for our estimate of $\bar{x}$:

$$CI_{95\%} = (\bar{x} - 1.96\sigma_{\bar{x}}, \bar{x} + 1.96\sigma_{\bar{x}})$$

or

$$CI_{95\%} = (\bar{x} - 1.96\frac{\sigma}{\sqrt{N}}), \bar{x} - 1.96\frac{\sigma}{\sqrt{N}}))$$

Warning

- Confidence does not mean probability!
- The confidence interval is the range of values that you expect your estimate to fall between a certain percentage of the time if you run your experiment infinitely many times or re-sample the population in the same way.
- If we were to repeat the process many times, (1-α)100% of the time our confidence interval captures the true parameter.

Reflection Questions

1. What are type I error, type II error, and power?
2. What is a p-value?

Summary

1. Sample statistics are random variables.
2. The behavior of a statistic is described by its samp*ling* distribution.
3. Under certain conditions (often reasonable), we can make assumptions about the properties of the samp*ling* distribution.
4. The statistic calculated from your sample is a realized value.
5. We can use knowledge of the sampling distribution to compare our sample/sample statistic to the population. It is thus possible to assess how improbable our sample (say, sample mean) is and make a decision.
6. Hypothesis testing is asymmetrically informative.
7. We can use our sample statistics as estimates of the parameters of the population it came from.

Next time

- Topic: What is regression?
- Read:
  - Neter Chapter 1 (to prepare)
  - Kleinbaum 5.1-5.6

References

These slides utilizes examples from:

https://st47s.com/Math158/Notes/intro.html#statistics-a-review