Review of Statistical Inference

Stats 102A

Miles Chen

Department of Statistics

Week 8 Friday

UCLA

Section 1

Review of Intro Stats

Descriptive Statistics - Measures of Center

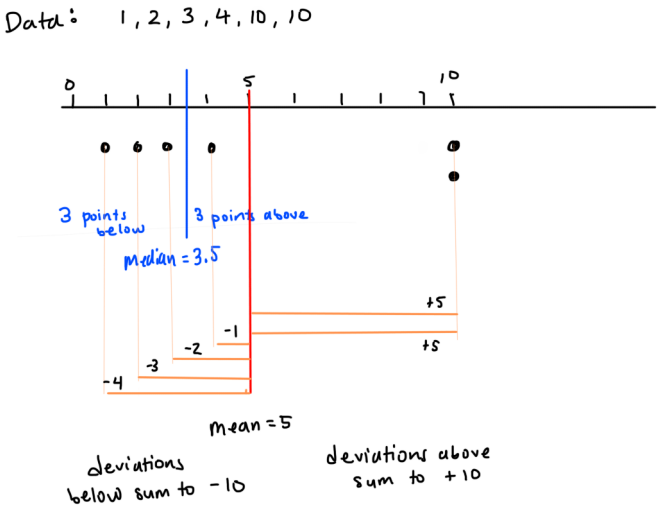
We can describe the center of a sample of data with the **mean** and the **median**.

Key conceptual summaries:

- The **median** balances the *number* of points that are lower or higher.
- The **mean** balances the *deviations* of points that are lower or higher.

Thus the mean will be 'pulled' in the direction of the skew. In general, the median is preferred for skewed distributions.

Descriptive Statistics - Measures of Center



Descriptive Statistics - Spread

Variance and **standard deviation** measure the spread of values.

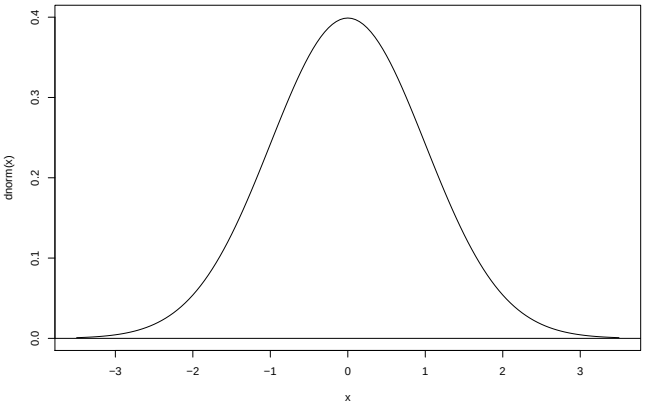
Smaller values of variance and standard deviation mean the values are closer together. (Ages of students in a 2nd grade classroom will have small variance - nearly everyone has the same age.)

Larger values of variance and standard deviation mean the data is more spread out. (Ages of people waiting at the DMV will have large variance - everyone has different ages.)

The variance is the "average" squared deviation from the mean. The standard deviation is the square root of the variance.

Key Ideas - Normal Distribution

The normal distribution is unimodal and symmetric. If data follows a normal distribution, most of the values are near the middle.



Normal Distribution - Empirical rule

68 is within ± 1 Standard Deviation of the mean

95 is within ± 2 Standard Deviation of the mean. (more accurately 1.96 sd)

99.7 is within ± 3 Standard Deviation of the mean.

Key Ideas - Populations and Samples

Every research question has an associated **population**.

The population is *everyone* who is relevant to answering the research question. The population is too big to be studied directly. We may not even know how to identify everyone in the population.

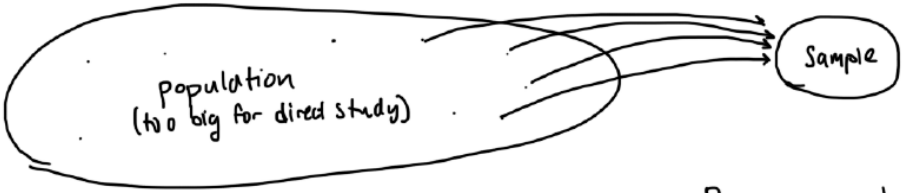
Example

- Research question: "Is experimental drug XYZ effective at treating disease JKL?"
- Population: Everyone who has disease JKL.

Key Ideas - Populations and Samples

In statistics, we select individuals from the population to be part of our sample.

We study the sample and based on what we see in the sample we make conclusions about the population.



Ideally, our sample is representative of the population. One method that should produce representative samples is **random sampling**. Keep in mind that random sampling is not guaranteed to produce representative samples.

Key Ideas - Sampling Distributions

When we take a random sample of data from the population, we often calculate summary statistics like the mean of the sample.

Because the sampling process is random, if we were to take another random sample, we would get different values and thus a different sample mean.

Some values of the sample mean are more likely and some values will be less likely.

The distribution of all possible values of the sample mean from all possible random samples is known as **the sampling distribution of the mean**.

Key Ideas - Sampling Distributions

We can create a sampling distribution via simulation.

1. We begin with a hypothesized population. We arbitrarily specify that the population has a certain shape (let's say normal) and that it has a certain mean (let's say 0) and a certain standard deviation (let's say 1).
2. We draw a random sample from this population and calculate the mean of the sample.
3. We draw another random sample from the population and calculate the mean of the second sample.
4. We can repeat this many times and record the sample mean for each random sample.
5. We look at the distribution of the sample means we produced.

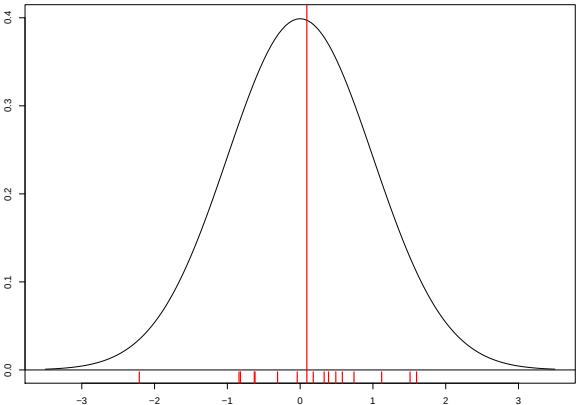
Central Limit Theorem

If the sample size is large enough, then the sampling distribution of the mean will follow a Normal distribution, with a standard deviation $= \sigma/\sqrt{n}$.

One random sample drawn from a Normal Population

```
##  [1] -2.21 -0.84 -0.82 -0.63 -0.62 -0.31 -0.04  0.18  0.33  0.39  0.49  0.58  0.74  1.12  1.51  1.60
```

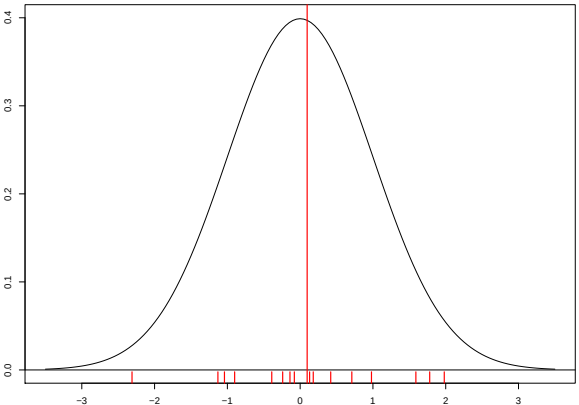
```
## Mean: 0.092
```



Another random sample drawn from a Normal Population

```
##  [1] -2.31 -1.13 -1.04 -0.90 -0.39 -0.24 -0.14 -0.08  0.13  0.18  0.42  0.71  0.98  1.59  1.78  1.98
```

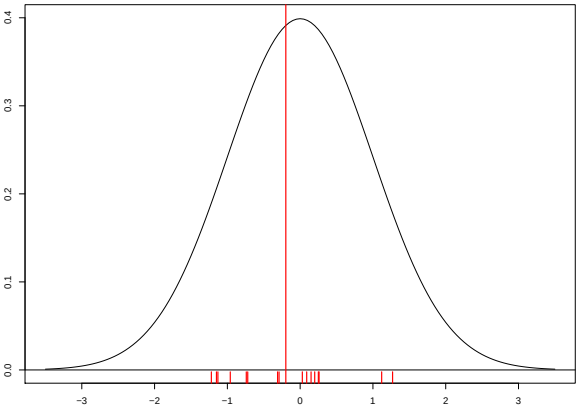
```
## Mean: 0.096
```



Another random sample drawn from a Normal Population

```
##  [1] -1.22 -1.15 -1.13 -0.96 -0.74 -0.72 -0.31 -0.29  0.03  0.09  0.15  0.20  0.25  0.26  1.12  1.27
```

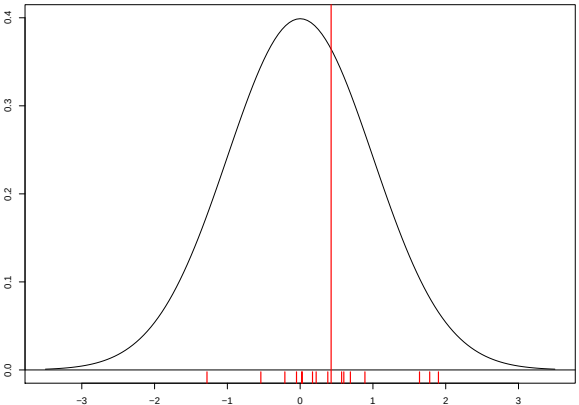
```
## Mean: -0.197
```



Another random sample drawn from a Normal Population

```
##  [1] -1.28 -0.54 -0.21 -0.05  0.02  0.03  0.17  0.22  0.38  0.57  0.60  0.69  0.89  1.64  1.78  1.90
```

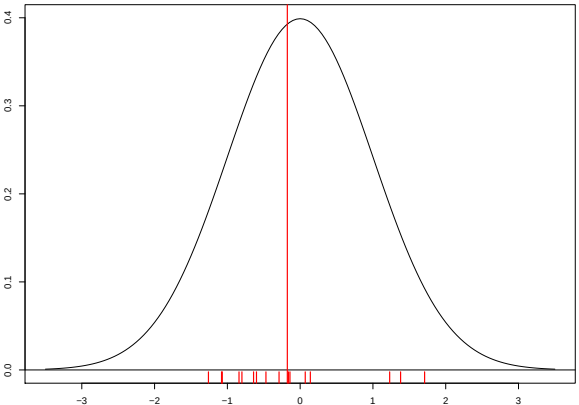
```
## Mean: 0.426
```



Another random sample drawn from a Normal Population

```
##  [1] -1.26 -1.08 -1.07 -0.84 -0.80 -0.64 -0.60 -0.47 -0.29 -0.16 -0.14  0.07  0.14  1.23  1.38  1.71
```

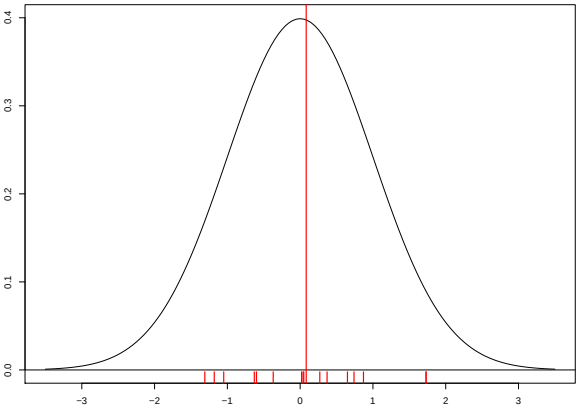
```
## Mean: -0.176
```



Another random sample drawn from a Normal Population

```
##  [1] -1.31 -1.18 -1.05 -0.63 -0.60 -0.37  0.02  0.04  0.05  0.27  0.37  0.65  0.74  0.87  1.73  1.73
```

```
## Mean: 0.083
```



Approximating a sampling distribution with many samples

I ran the code to generate 10,000 different random samples of 16 drawn from the standard normal distribution. For each sample, I calculated the sample mean and stored it in the vector `sample_means`

```
sample_means <- rep(NA, 10 ^ 4)
for(i in seq_along(sample_means)){
  rand_samp <- rnorm(16)
  sample_means[i] <- mean(rand_samp)
}
mean(sample_means) # very close to 0
```

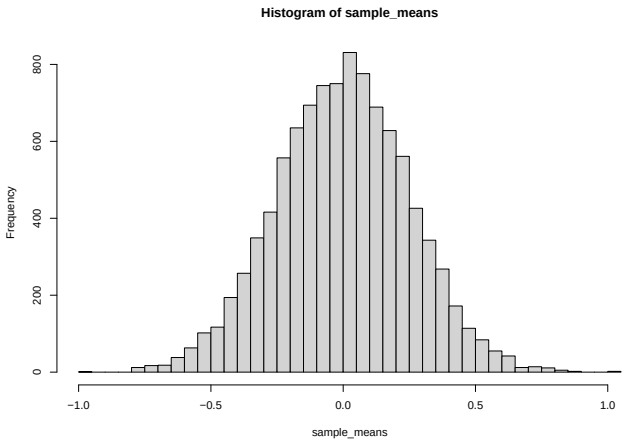
```
## [1] -0.001504151
```

```
sd(sample_means)   # very close to sigma / sqrt(16)
```

```
## [1] 0.2501872
```

Histogram of the Sampling Distribution of the Mean

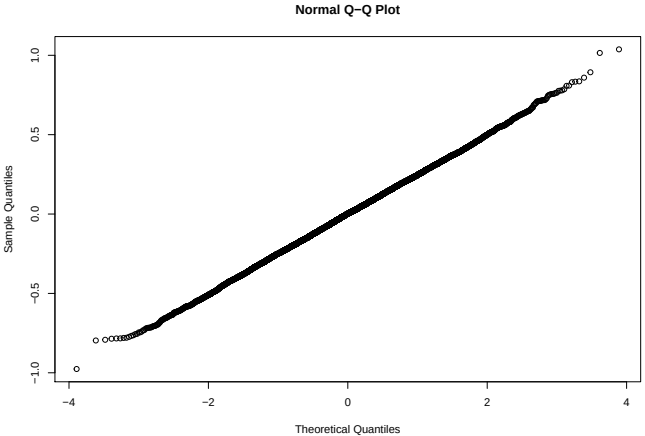
```
hist(sample_means, breaks = 30)
```



qq plot

The qq plot shows that the distribution of sample means follow the normal distribution quite closely.

```
qqnorm(sample_means)
```



Key Ideas - Statistical Inference

Provided that the sample is large enough:

The Central Limit Theorem establishes that the mean of a random sample comes from the normal distribution centered at μ (the population mean) with standard deviation $\sigma/\sqrt{n}$

The empirical rule tells us that 95% of the time, a random value drawn from a normal distribution will be within 2 standard deviations of the mean.

Thus: When I draw a random sample, there is a 95% probability that the mean of the random sample will be within 2 standard deviations of the population mean.

Key Ideas - Confidence Interval

Armed with this knowledge, when we observe a random sample with a mean $\bar{x}$, we can be 95% confident that the population mean μ is within 2 standard deviations ($\sigma/\sqrt{n}$) of $\bar{x}$.

This forms the foundation of a confidence interval. The only problem is that the exact standard deviation $\sigma/\sqrt{n}$ is unknown, so we estimate it using $s/\sqrt{n}$. This extra uncertainty means we must use the t-distribution instead of the normal distribution to determine the critical value. We won't use 1.96 but a number slightly larger.

$$\bar{x} \pm t \times \frac{s}{\sqrt{n}}$$

Hypothesis Testing

We can approach statistical inference from another approach:

If we imagine a hypothetical population with a hypothetical mean and hypothetical standard deviation, we can perform calculations to see how likely such a population could produce a random sample that has the properties of the observed sample that we have.

Hypothesis Testing - the t-test

The **two independent sample t-test** takes samples from two populations and tests to see if the respective populations have the same mean or not.

The Null Hypothesis

The null hypothesis states that there is no significant difference between the means of the specified populations. Any difference we observe between the means of our samples is due to sampling or experimental error.

The p-value

The p-value is the probability of observing a summary statistic equal to or more extreme than what was actually observed when the null hypothesis is true.

t-test - The big idea

We imagine two hypothetical populations that have the same mean.

When we take a random sample from each population, *we expect both samples to have sample means that are approximately equal*. (We expect the sample means to have a difference of zero.)

If they are not equal, we want to know if the observed difference between them is significantly different from zero. $(\bar{x}_1 - \bar{x}_2)$

We answer this by finding the p-value. If the two populations were equal, what is the probability that from random sampling we would get two samples where the difference between sample means is equal to or greater than the observed difference?

Null Hypothesis: population 1 and population 2 have the same mean value

$$H_0 : \mu_1 = \mu_2$$

Let's say you collect data, and that $\bar{x}_1 = 10$ and $\bar{x}_2 = 12$.

Is the observed difference of 2 significant?

As is always the case, the answer depends. In this example, it depends on the standard error.

One scenario: the standard error is very small, so the observed difference of 2 is much larger than (say 5 times) the SE. In this case, the difference is significant.

Another scenario: the standard error is large, so the observed difference of 2 is NOT much larger than (say 1.2 times) the SE. In this case, the difference is NOT significant.

p-values

Our use of p-values simply quantifies the probability of observing our data (or something more extreme) when the null hypothesis is true (observing our data from random chance alone). We compare our p-value to an arbitrarily selected significance level (alpha, α) to decide if the p-value is small or large.

So if our **p-value is small**, it means that our observed difference is so large, that it is unlikely to see these results if the only source of variation is random chance. Thus, we can be fairly confident to conclude that our data is not a result of random chance alone, but rather that the alternative hypothesis is true. **We reject the null hypothesis**.

On the other hand, if the **p-value is large**, it means that our observed difference is still quite probable to occur even when the only source of variation is random chance. So this does not prove that the null hypothesis is true, but that we cannot eliminate the possibility that our data is a result of random chance. **We do not reject the null hypothesis**.

Parametric Hypothesis Tests

In your intro stats class, your hypothesis tests probably used the normal distribution (when you dealt with proportions) or the t-distribution (when you dealt with means). These relied on the underlying mathematics of the Central Limit Theorem (proved in your Stats 100B class), which showed that the sampling distributions of the sample mean or sample proportion approached the normal distribution.

Of course, you had to check a set of conditions to make sure that it was okay to use the Central Limit Theorem

Next week: Randomization Tests

In a randomization test, we will test the null hypothesis that there is no difference between group means.

A randomization test does not depend on the Central Limit Theorem, but rather, actual randomization.

Randomization tests depend on a fundamental assumption of **exchangeability**.

Bur first, a review on data collection

Section 2

Concept Review: Data Collection

How we got the data matters

Suppose a dentist wants to know if a daily dose of 500 mg of vitamin C will result in fewer canker sores in the mouth than taking no vitamin C.

Scenario 1

The dentist, working though the local dental society, convinces all of the dental patients in town **with appointments in the first two weeks in December** to be subjects of a study.

He divides them into two groups, **those who normally take at least 500 mg of vitamin C each day** and **those who do not**.

He asks them how often they have canker sores in their mouths and checks their dental record to see who has complained of canker sores.

He compares the proportion of those who take vitamin C and complain about canker sores with the proportion of those who don't take vitamin C daily and complain of canker sores.

There is a significantly difference in the two proportions, with a significantly smaller proportion of those taking vitamin C having canker sores.

What can we conclude?

Since the patients do not represent a random sample from any population, it is not possible to make any inference about this result holding for a larger population.

Since the study was observational, with subjects not randomly assigned to treatments, no causal inference can be made.

We just know that for these patients, those who take vitamin C have fewer canker sores than those who don't. We don't know why, and we don't know if this result would be consistent with another group.

Scenario 2

The dentist, working though the local dental society, convinces all of the dental **patients with appointments in the first two weeks in December** to be subjects of a study.

He **randomly assigns** half of them to take 500 mg of vitamin C each day and half to abstain from taking Vitamin C for three months.

At the end of this time he determines the proportion of each group that has suffered from canker sores during those three months. There is a significant difference in the two proportions, with a significantly smaller proportion of those taking vitamin C having canker sores.

What can we conclude?

Since the patients do not represent a random sample from any population, it is not possible to make any inference that this result would hold for a larger population.

However, the treatments were randomly assigned to the subjects, so (assuming other factors were controlled or randomized) the difference in proportions having canker sores can be attributed to the vitamin C.

We don't know if this result would be consistent with another group, but we believe we know why, for this group, the proportions differ.

Scenario 3

The dentist, working though the local dental society, **selects a random sample** of the dental patients in town and convinces them to be subjects of a study.

He divides them into two groups, those who normally take at least 500 mg of vitamin C each day and those who do not.

Then he asks them how often they have canker sores in their mouths and then checks their dental record to see who has complained of canker sores. He compares the proportion of those who take vitamin C and complain of canker sores with the proportion of those who do not take vitamin C and complain of canker sores.

There is a significantly difference in the two proportions, with a significantly smaller proportion of those taking vitamin C having canker sores.

What can we conclude?

Since the patients selected were a random sample of dental patients in town, we can infer that the results observed in this study would be consistent with results from the whole population of dental patients in this town.

However, since the study was observational, with subjects not being randomly assigned to treatments, no causal inference can be made. For the population of dental patients in this town, we know that those taking vitamin C have fewer canker sores than those who didn't. We don't know if it is the vitamin C that causes this reduction or some other confounding variable.

We cannot conclude that for the general population, those taking vitamin C have fewer canker sores, since the sample was only of dental patients.

If the dental patients in this town are representative of dental patients in general, we can infer that dental patients who take vitamin C tend to have fewer canker sores than those who don't.

Scenario 4

The dentist, working though the local dental society, selects a **random sample** of dental patients in town and convinces them to be subjects of a study.

He **randomly assigns** half of them to take 500 mg of vitamin C daily for three months.

At the end of this time he determines that proportion of each group that has suffered canker sores during those three month.

There is a significant difference in the two proportions with a significantly smaller proportion of those taking vitamin C having canker sores.

What can we conclude?

Since the patients selected were a random sample of dental patients in town, we can infer than the results observed in this experiment would be consistent with results from the whole population of dental patients in this town.

Moreover, the treatments were randomly assigned to the subjects, so (assuming other factors were controlled or randomized) the difference in proportions having canker sores can be attributed to the vitamin C.

For the population of dental patients in this town, we know that those taking vitamin C have fewer canker sores than those who don't. Also, we believe that the reduction in canker sores is a consequence of taking the vitamin C.

We cannot conclude that for the general population, those taking vitamin C have fewer canker sores, since the sample was only of dental patients.

If the dental patients in this town are representative of dental patients in general, we can infer that dental patients who take vitamin C tend to have fewer canker sores than those who don't, as a result of taking the vitamin C.

Reality

Scenario 4 - the random sample with random assignment is a nice fantasy for statisticians.

The reality is that having both a random sample and random assignment is almost never possible.

Experiments require patient/participant consent, and as soon as one person declines participation, the group is no longer a random sample.

Finding participants and gathering data from an experiment is an arduous process.