THE UNIVERSITY OF TEXAS AT AUSTIN
McCOMBS SCHOOL OF BUSINESS

Review of distributions and estimation

Lecture 8

STA 371G

The story of statistics

- Usually, there is an *population* that we are interested in—and a *parameter* about that population we care about—but we don't have access to the population and all we have is a *sample*

The story of statistics

- Usually, there is an *population* that we are interested in—and a *parameter* about that population we care about—but we don't have access to the population and all we have is a *sample*
- So we calculate a *statistic* in the sample and use that as our best estimate of the population parameter

The story of statistics

- Usually, there is an *population* that we are interested in—and a *parameter* about that population we care about—but we don't have access to the population and all we have is a *sample*
- So we calculate a *statistic* in the sample and use that as our best estimate of the population parameter
- Example: I want to know the average GPA at UT, but I only have a sample of $n = 100$ GPAs.

The story of statistics

- Usually, there is an *population* that we are interested in—and a *parameter* about that population we care about—but we don't have access to the population and all we have is a *sample*
- So we calculate a *statistic* in the sample and use that as our best estimate of the population parameter
- Example: I want to know the average GPA at UT, but I only have a sample of $n = 100$ GPAs.

Population	All GPAs at UT
Sample	The 100 GPAs in my sample
Parameter	Average GPA among all UT students (μ)
Statistic	Average GPA among the 100 students in my sample ($\hat{\mu}$)

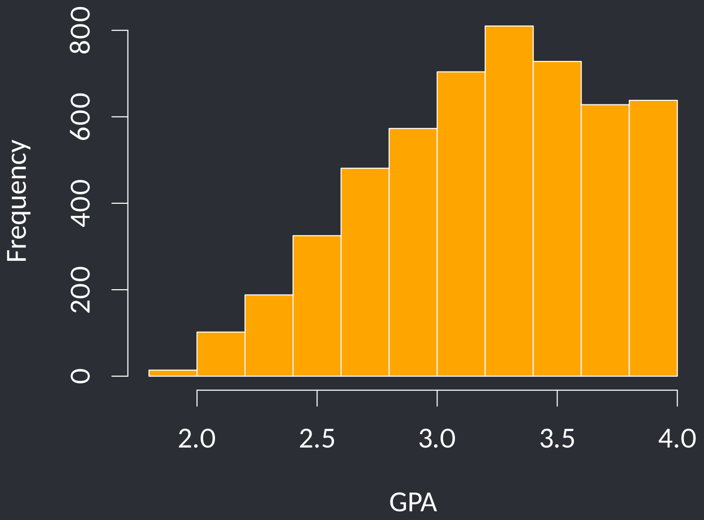
Data set

The data set `ut2000` contains information on all 5191 students that entered UT Austin in Fall 2000 and graduated within 6 years.

```
head(ut2000)
```

```
  SAT.V SAT.Q SAT.C          School  GPA Status
1   690   580  1270        BUSINESS 3.82      G
2   530   710  1240 NATURAL SCIENCE 3.53      G
3   610   700  1310 NATURAL SCIENCE 3.37      G
4   730   700  1430     ENGINEERING 3.34      G
5   700   710  1410 NATURAL SCIENCE 3.72      G
6   540   690  1230    LIBERAL ARTS 2.69      G
```

```r
hist(ut2000$GPA, main="", col="orange", xlab="GPA")
```



Let's take a sample

Usually, we only have access to a sample of the data. Let's pretend that we only had a sample of $n = 100$ students:

```
sample.gpas <- sample(ut2000$GPA, 100)
mean(sample.gpas)

[1] 3.23
```

Since we have a random sample, it's a good, but not perfect, estimate of the population GPA (3.212).

Let's take a sample

Usually, we only have access to a sample of the data. Let's pretend that we only had a sample of $n = 100$ students:

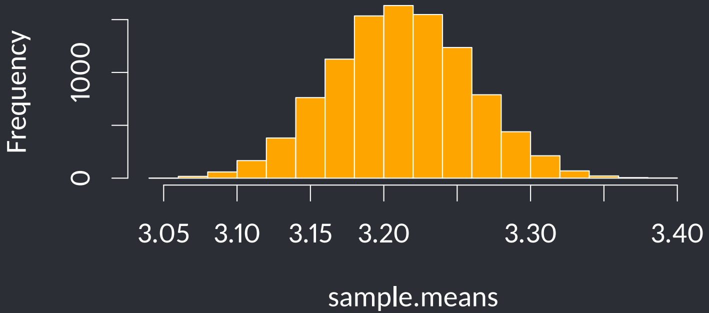
```
sample.gpas <- sample(ut2000$GPA, 100)
mean(sample.gpas)

[1] 3.23
```

Since we have a random sample, it's a good, but not perfect, estimate of the population GPA (3.212). But normally we don't have access to the population, so we don't know how good our estimate is!

Normally we only have access to one sample. But what if we had many samples, and we took the sample mean in each sample?

```
sample.means <- replicate(10000,
  mean(sample(ut2000$GPA, 100)))
hist(sample.means, main="", col="orange")
```



Sampling distribution of $\overline{GPA}$

The *sampling distribution* of $\overline{GPA}$ is the distribution of sample means, if we took repeated samples:

$$E(\overline{GPA}) = \mu = 3.212$$
$$\mathrm{SD}(\overline{GPA}) = \frac{\sigma}{\sqrt{n}} = \frac{0.48}{\sqrt{100}} = 0.048$$

The last value quantifies how much the sample mean will vary from sample to sample. But we normally can't compute σ since we don't have the whole population, so we estimate it by calculating the SD in the *sample* ($\hat{\sigma}$) and dividing by $\sqrt{n}$; this is the *standard error of the mean*.

Confidence intervals

Given a sample mean, we want to calculate a *confidence interval*, which gives a plausible range of values for the population mean.

Confidence intervals

Given a sample mean, we want to calculate a *confidence interval*, which gives a plausible range of values for the population mean. A confidence interval is always of the form

$$\text{sample statistic} \pm (\text{critical value})(\text{standard error}).$$

Confidence intervals

Given a sample mean, we want to calculate a *confidence interval*, which gives a plausible range of values for the population mean. A confidence interval is always of the form

$$\text{sample statistic} \pm (\text{critical value})(\text{standard error}).$$

Our sample statistic is $\hat{\mu}$ and our standard error is $\hat{\sigma}/\sqrt{n}$.

Confidence intervals

Given a sample mean, we want to calculate a *confidence interval*, which gives a plausible range of values for the population mean. A confidence interval is always of the form

$$\text{sample statistic} \pm (\text{critical value})(\text{standard error}).$$

Our sample statistic is $\hat{\mu}$ and our standard error is $\hat{\sigma}/\sqrt{n}$. What is the critical value?

As it turns out, the sampling distribution (of $\hat{\mu}$) is not *quite* Normal. If we standardize the sample means, the distribution of

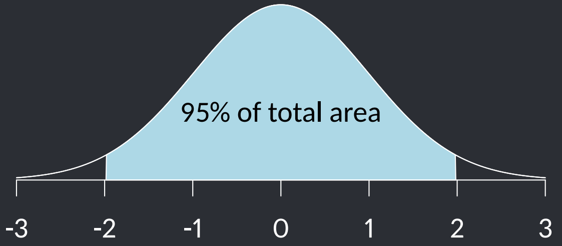
$$\frac{\hat{\mu} - \mu}{\hat{\sigma}/\sqrt{n}}$$

is called a t-distribution with $n - 1$ degrees of freedom.

As it turns out, the sampling distribution (of $\hat{\mu}$) is not *quite* Normal. If we standardize the sample means, the distribution of

$$\frac{\hat{\mu} - \mu}{\hat{\sigma}/\sqrt{n}}$$

is called a *t*-distribution with $n - 1$ degrees of freedom. The critical value for a 95% confidence interval is $t^* = \pm 1.984$, the value that cuts off 95% of the area under the *t*-distribution:



There are two helpful R functions for calculating values around *t*-distributions:

- `pt(x, df)` calculates $P(t < x)$ if we are looking at a distribution with *df* degrees of freedom.

There are two helpful R functions for calculating values around t-distributions:

- `pt(x, df)` calculates $P(t < x)$ if we are looking at a distribution with *df* degrees of freedom.

There are two helpful R functions for calculating values around t-distributions:

- `pt(x, df)` calculates $P(t < x)$ if we are looking at a distribution with df degrees of freedom.
- `qt(y, df)` does the opposite; it figures out what value of x will give $P(t < x) = y$.

There are two helpful R functions for calculating values around t-distributions:

- `pt(x, df)` calculates $P(t < x)$ if we are looking at a distribution with *df* degrees of freedom.
- `qt(y, df)` does the opposite; it figures out what value of x will give $P(t < x) = y$.

There are two helpful R functions for calculating values around t-distributions:

- `pt(x, df)` calculates $P(t < x)$ if we are looking at a distribution with df degrees of freedom.
- `qt(y, df)` does the opposite; it figures out what value of x will give $P(t < x) = y$.

So the critical value for a 95% confidence interval is `qt(.975, 99)` when $n = 100$.

There are two helpful R functions for calculating values around t-distributions:

- `pt(x, df)` calculates $P(t < x)$ if we are looking at a distribution with df degrees of freedom.
- `qt(y, df)` does the opposite; it figures out what value of x will give $P(t < x) = y$.

So the critical value for a 95% confidence interval is `qt(.975, 99)` when $n = 100$.

There are similar functions `pnorm` and `qnorm` when you are working with Normal distributions.

Using our sample of $n = 100$, we calculate that a 95% confidence interval for the mean population GPA is

$$\hat{\mu} \pm t^* \frac{\hat{\sigma}}{\sqrt{n}}$$

Using our sample of $n = 100$, we calculate that a 95% confidence interval for the mean population GPA is

$$\hat{\mu} \pm t^* \frac{\hat{\sigma}}{\sqrt{n}} \quad \Longrightarrow \quad 3.226 \pm 1.984 \cdot \frac{0.471}{\sqrt{100}}$$

Using our sample of $n = 100$, we calculate that a 95% confidence interval for the mean population GPA is

$$\hat{\mu} \pm t^* \frac{\hat{\sigma}}{\sqrt{n}} \quad \Longrightarrow \quad 3.226 \pm 1.984 \cdot \frac{0.471}{\sqrt{100}} \quad \Longrightarrow \quad (3.133, 3.32)$$

Using our sample of $n = 100$, we calculate that a 95% confidence interval for the mean population GPA is

$$\hat{\mu} \pm t^* \frac{\hat{\sigma}}{\sqrt{n}} \implies 3.226 \pm 1.984 \cdot \frac{0.471}{\sqrt{100}} \implies (3.133, 3.32)$$

There are two ways to interpret this:

- **Informally,** we are 95% confident that the population mean GPA is between 3.133 and 3.32.

Using our sample of $n = 100$, we calculate that a 95% confidence interval for the mean population GPA is

$$\hat{\mu} \pm t^* \frac{\hat{\sigma}}{\sqrt{n}} \quad \Longrightarrow \quad 3.226 \pm 1.984 \cdot \frac{0.471}{\sqrt{100}} \quad \Longrightarrow \quad (3.133, 3.32)$$

There are two ways to interpret this:

- **Informally,** we are 95% confident that the population mean GPA is between 3.133 and 3.32.
- **Formally,** if we took repeated samples and found the 95% CI within each sample, 95% of the CIs would contain the population mean.

R can do this work for you!

```r
t.test(sample.gpas, conf.level=0.95)
```

```


One Sample t-test


data:  sample.gpas
t = 70, df = 100, p-value <2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 3.13 3.32
sample estimates:
mean of x
     3.23
```

Hypothesis tests

- Let's say Mr. Sooner comes along and claims that the average UT GPA is actually 3.0.

Hypothesis tests

- Let's say Mr. Sooner comes along and claims that the average UT GPA is actually 3.0.
- Usually, we don't have the population, and so we can't know for sure that he is wrong.

Hypothesis tests

- Let’s say Mr. Sooner comes along and claims that the average UT GPA is actually 3.0.
- Usually, we don’t have the population, and so we can’t know for sure that he is wrong.
- But we do have some evidence (our sample) that we can bring to bear on the question.

Let's start by framing Sooner's claim as a null hypothesis; the alternative hypothesis is what we will believe if it turns out the null is false:

$$
\begin{array}{ll|l}
H_0 & \text{(null hypothesis)} & \mu = 3.0 \\
H_A & \text{(alternative hypothesis)} & \mu \neq 3.0
\end{array}
$$

Let's start by framing Sooner's claim as a null hypothesis; the alternative hypothesis is what we will believe if it turns out the null is false:

$$\begin{array}{lll|l} H_0 & \text{(null hypothesis)} & & \mu = 3.0 \\ H_A & \text{(alternative hypothesis)} & & \mu \neq 3.0 \end{array}$$

A hypothesis test will produce a *p*-value, which represents the probability:

$$P(\text{seeing data this extreme in our sample} \mid H_0 \text{ is true})$$

Let’s start by framing Sooner’s claim as a null hypothesis; the alternative hypothesis is what we will believe if it turns out the null is false:

$$\begin{array}{lll|l} H_0 & \text{(null hypothesis)} & & \mu = 3.0 \\ H_A & \text{(alternative hypothesis)} & & \mu \neq 3.0 \end{array}$$

A hypothesis test will produce a *p*-value, which represents the probability:

$$P(\text{seeing data this extreme in our sample} \mid H_0 \text{ is true})$$

In other words, if Sooner is correct, how likely is it that in our sample we would see a sample mean that is so far away from his hypothesized value?

R can run hypothesis tests for us:

```
t.test(sample.gpas, mu=3)



One Sample t-test


data:  sample.gpas
t = 5, df = 100, p-value = 5e-06
alternative hypothesis: true mean is not equal to 3
95 percent confidence interval:
 3.13 3.32
sample estimates:
mean of x
     3.23
```

So:

$$P(\text{seeing data this extreme in our sample} \mid H_0 \text{ is true}) = 5 \times 10^{-6}.$$

So:

$$P(\text{seeing data this extreme in our sample} \mid H_0 \text{ is true}) = 5 \times 10^{-6}.$$

Since we *did* see data this extreme in our sample, it suggests that we should *reject* the null hypothesis as implausible. (Sorry, Mr. Sooner!)

So:

$$P(\text{seeing data this extreme in our sample} \mid H_0 \text{ is true}) = 5 \times 10^{-6}.$$

Since we *did* see data this extreme in our sample, it suggests that we should *reject* the null hypothesis as implausible. (Sorry, Mr. Sooner!)

When doing hypothesis testing, we select an α value a priori and then reject the null hypothesis if $p < \alpha$.

So:

$P(\text{seeing data this extreme in our sample} \mid H_0 \text{ is true}) = 5 \times 10^{-6}$.

Since we *did* see data this extreme in our sample, it suggests that we should *reject* the null hypothesis as implausible. (Sorry, Mr. Sooner!)

When doing hypothesis testing, we select an α value a priori and then reject the null hypothesis if $p < \alpha$.

$\alpha = .05$ is a good "default" to use unless you have a reason to set it higher or lower.