Statistics Review

AGEC 317: Economic Analysis for Agribusiness Management

Instructor: Michael Black

TEXAS A&M UNIVERSITY
Agricultural
Economics

Statistics

Why should you care about statistics?

Statistics

Consider the following scenario:
You own a butcher shop, and you can purchase one of two machines that make hamburger patties:

- Machine A: 1600 patties/hour $\pm$ 200, with each patty at 4oz., $\pm$ 0.03oz.
- Machine B: 1550 patties/hour $\pm$ 25, with each patty at 4oz., $\pm$ 0.01oz.

Which machine should you purchase?

Statistics

Your professional life will be **full** of decisions like this, where money is often on the line. How do you make your decision?

- Ask your friends?
- Read customer reviews?
- The cheapest sticker price?

The real world is messy and uncertain. Statistics and mathematics allow us to make informed rational decisions in the face of uncertainty.

Statistics

With statistics, we can

- Present and describe data
- Estimate functions (demand, production, cost equations, etc)
- Make inferences about a population using a random sample
- Forecast
- ...and much more

Data

We start with *data*. Without data, the modern economist is increasingly obsolete. It is a valuable asset, and to navigate your future job you will need to know how to correctly handle data. Let's start with some key definitions.

Data

- **Population**: the broad collection of units (people) that we make inferences about. It is almost always the case that we can't observe the entire population, hence the need for statistics!
- **Sample**: a subset of the population on whom we perform statistical analysis.

Ex.: We often infer the percent of Americans who support/don't support a political figure. What is the statistical population? How would you gather the sample?

Data

- **Variable**: a symbolic number that can take many different values (get it? *vari*-able?)

Ex.:

$$y = 4x + 6$$

What are the variables of the function?

Data

- **Parameter** (or **coefficient**): A number in front of a variable, that defines the behavior of that variable on the overall function.

Ex.:

$$y = 4x + 6$$

What are the parameters (coefficients) of the function? What do they mean?

Data

- **Random**: We say something is random when it exhibits some unpredictable or messy or imperfect pattern. Variables are often random because they can take many values and their exact value is not always predictable. However, even random events follow some probability distribution.

Ex.: Next time you cook with eggs, throw one on the ground as hard as you can. The gooey mess will make a different pattern on your floor every time you do this. Random pattern, right? Yes, but the probability of egg getting on your floor is much higher than the probability of egg getting on your ceiling. Even though the event is random, there is still some probability distribution governing the event.

Data

Data is often organized as a matrix. The columns are variables, the rows are observations.

	A	B	C	D	E	F	G
1		Carat	Color	Clairity	Depth	PricePerCt	Total Price
2	1	1.08	"E"	"VS1"	68.6	6693.3	7228.8
3	2	0.31	"F"	"VVS1"	61.9	3159	979.3
4	3	0.31	"H"	"VS1"	62.1	1755	544.1
5	4	0.32	"F"	"VVS1"	60.8	3159	1010.9
6	5	0.33	"D"	"IF"	60.8	4758.8	1570.4
7	6	0.33	"G"	"VVS1"	61.5	2895.8	955.6
8	7	0.35	"F"	"VS1"	62.5	2457	860
9	8	0.35	"F"	"VS1"	62.3	2457	860
10	9	0.37	"F"	"VVS1"	61.4	3402	1258.7
11	10	0.38	"D"	"IF"	60	5062.5	1923.8
12	11	0.38	"E"	"VVS2"	61.5	3496.5	1328.7

Data

Variables come in different flavors

- Categorical: numbers (often 0 and 1) that indicate categories, like race and gender
- Numerical
  - Discrete: numbers that are usually integers (whole) that have meaning, like the number of pets you have
  - Continuous: numbers that can take any value, like the amount of oil from a well (18.902374923847... bbl/day)

Data

Where do we get data?

Classic Sources:

- USDA NASS Quick Stats
- USDA Economic Research Service
- Federal Reserve Economic Data
- Bureau of Economic Analysis

Evolving Sources:

- Reddit: Datasets
- FiveThirtyEight
- Kaggle

Data

...but data is everywhere. The hunt for data to answer a question is often a very important and time consuming part of any project.

Data

Broad types of data:

- Primary data: you collect yourself
- Secondary data: someone else collected for you

Broad ways to collect data:

- Experiment
- Survey
- Observe

Descriptive statistics

Once we find data, we want to describe and present key information. *Communication* will be a very important part of your job. The most advanced analysis in the world is meaningless if it is not communicated properly.

- Tables
- Figures

Important notation

Recall our definition of **random** and **variable**. A **random variable** is then straightforward: it is a variable that takes random individual observations. Remember that by "random" we don't mean "OMG like that was like SO random!". We observe statistical order in the chaos.

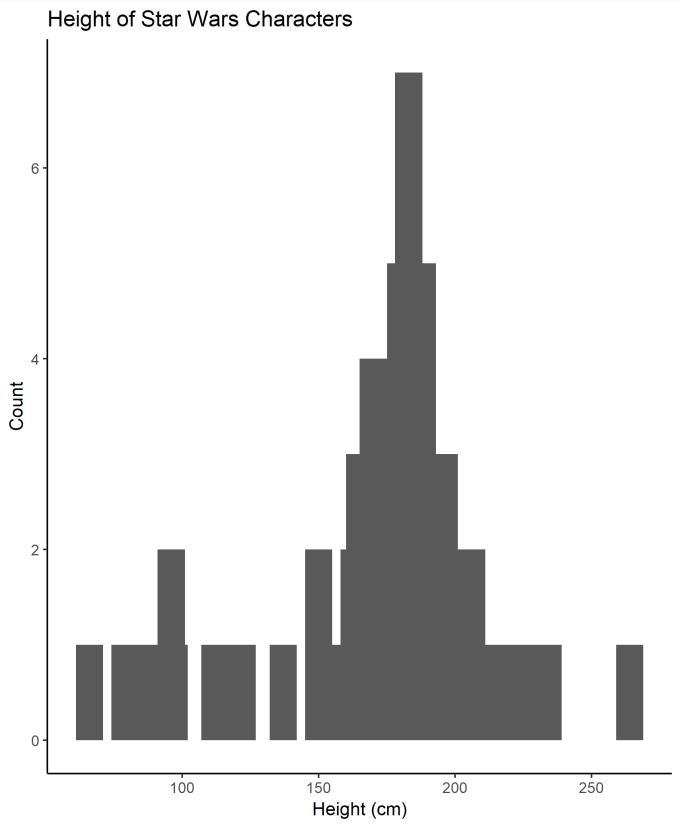
Important notation

In general, we denote a random variable with a capital letter like X, and an individual observation with a lower case letter and index like x_i.

$\Rightarrow$ The time you wake up each morning is a random variable, and the time you woke up this morning is an individual observation.

Histogram

A **histogram** describes the frequency of certain values for a variable. For example, here is a histogram of the height of Star Wars characters:



Histogram

To develop a histogram, we first need to determine the frequency distribution: or the count of certain values. For example, the data on Star Wars characters looks like the following:

	name	height
	<chr>	<int>
1	Luke Skywalker	172
2	C-3PO	167
3	R2-D2	96
4	Darth Vader	202
5	Leia Organa	150
6	Owen Lars	178
7	Beru Whitesun lars	165
8	R5-D4	97
9	Biggs Darklighter	183
10	Obi-Wan Kenobi	182

... with 77 more rows

Histogram

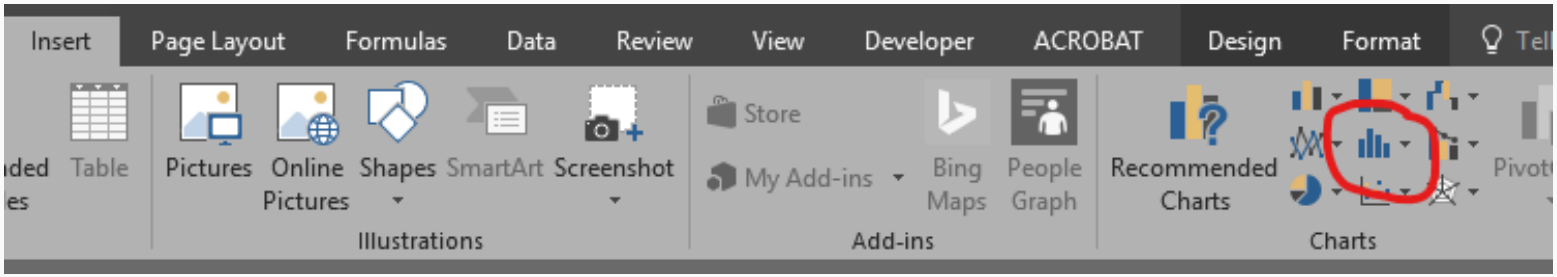
You can create a histogram by hand (the hard way), or use software (the easy way). By hand, the steps are:

1. Decide on the number of bins
2. Determine bin width
3. Decide where to cut off the chart (should you include outliers?)
4. Count the number of observations in each bin
5. Calculate the relative frequency of each bin

Of course, none of these are hard rules, so the final product will come with trial and error

Histogram

Alternatively, we can use Excel. The process for creating a histogram is now: highlight the data, and press the histogram button



By formatting the x-axis, (Right click $->$ Format Axis), you can adjust the number of bins or bin width until you are happy with the communication goal

Histogram

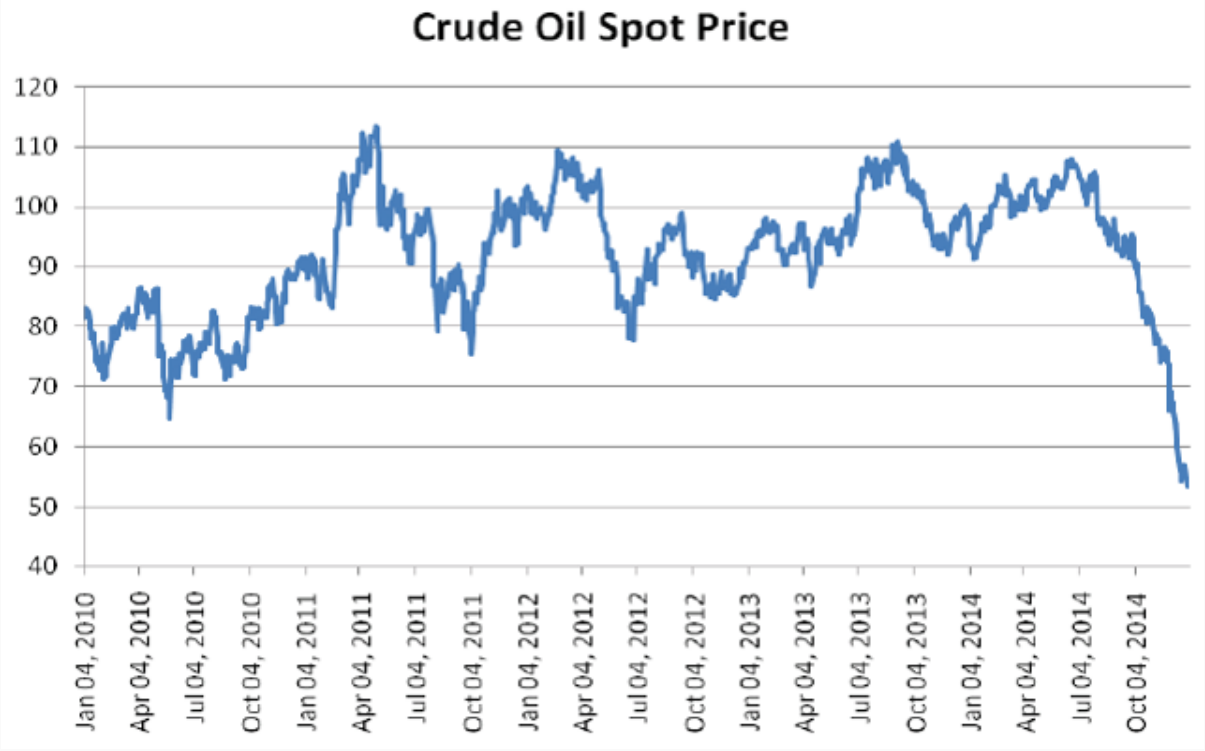
We can also use a programming language like R.

```
20  install.packages("ggplot2", "dplyr")
21  library(ggplot2, dplyr)
22  ggplot(starwars, aes(x = height))+
23    geom_bar(width = 10) +
24    theme_classic()+
25    labs(y = "Count", x = "Height (cm)", title = "Height of Star Wars Characters")
```

Excel requires lots of clicking buttons, while the R script can be reprodced exactly by anyone with the code.

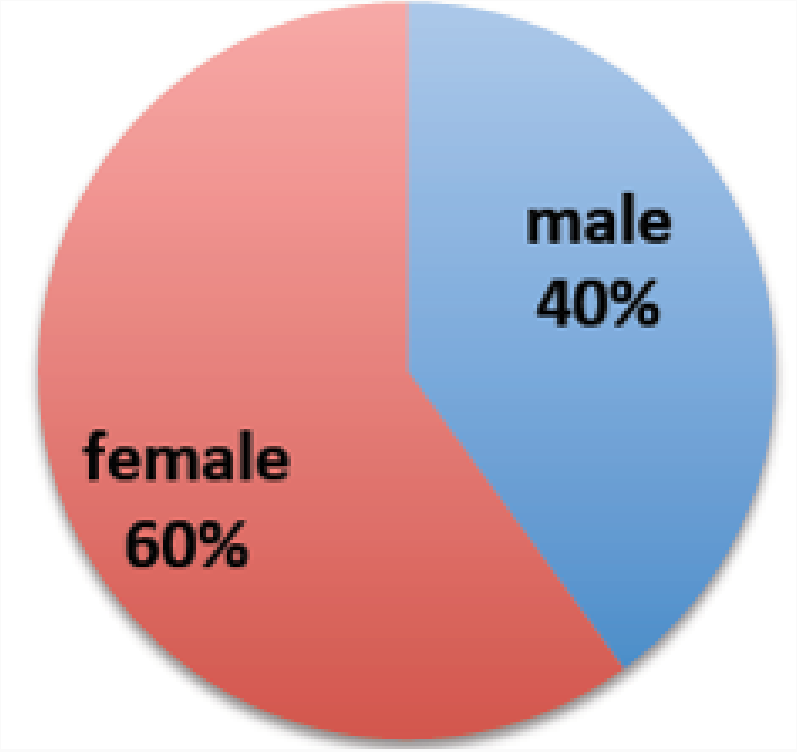
Line chart

A **line chart** is a powerful communication tool for data that changes over time. It is, as it sounds, a line that connects observations with a straight line. For time-series data, this can be very powerful.



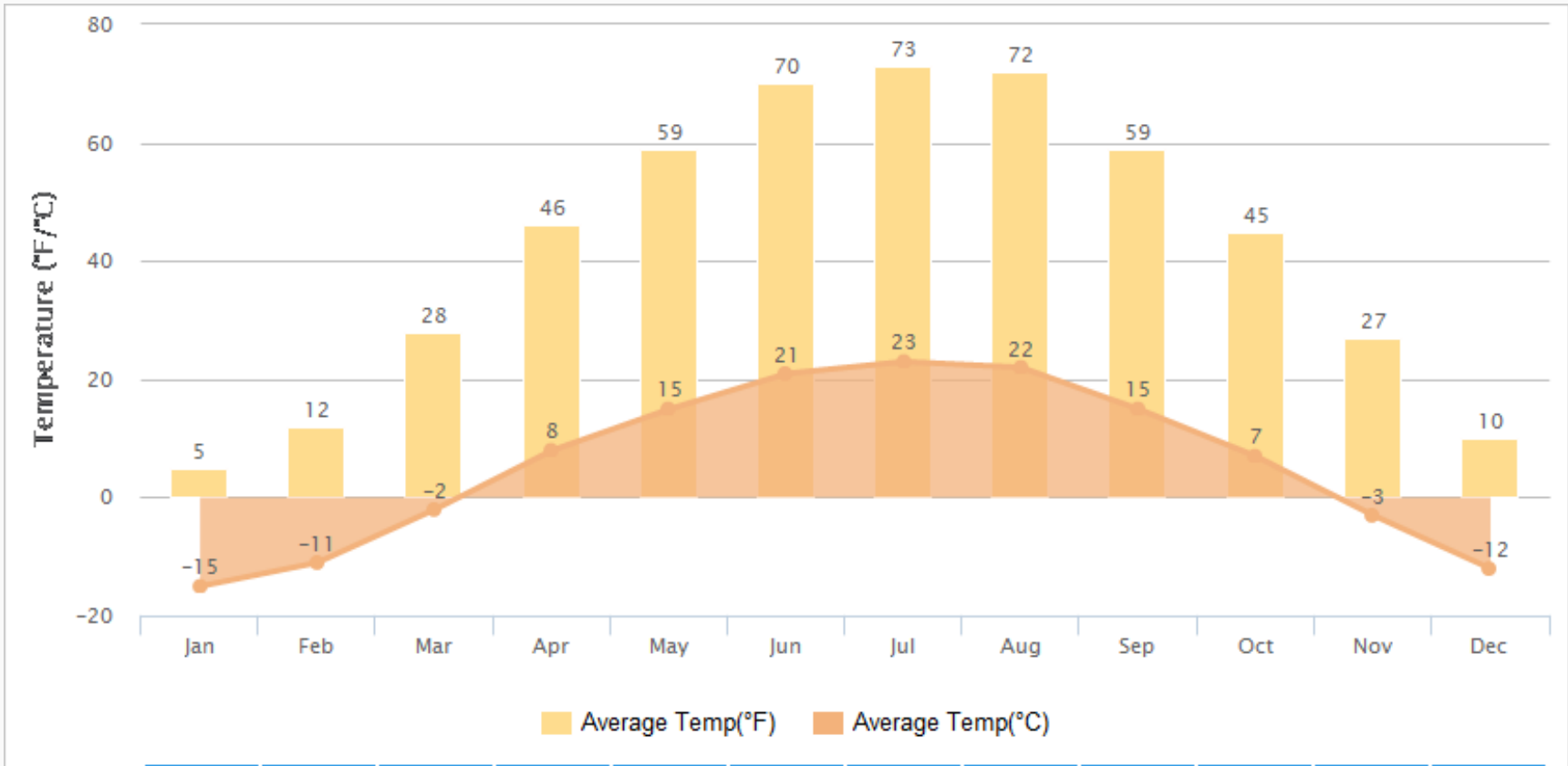
Pie chart

A **pie chart** is generally not great at communicating data. If your goal is to communicate differences in proportions, a bar graph is more effective unless the proportions are vastly different.



Bar graph

A **bar graph** uses bars to represent data, and is very common in summary statistics. A histogram is a specific example of a bar graph that is representing frequency of certain observations. A bar graph in general can represent many different types of data.



Describing data

Data can be described and summarized in several ways:

- Measures of location: mean, median, mode
- Measures of dispersal: range, variance, standard deviation
- Measures of shape: kurtosis

Mean

If there are N individuals in a population, and x_i is a single observation of some value, the **population mean** is:

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

And if there are n individuals in a sample (with $n < N$), then the **sample mean** is:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Median

The **median** of a variable is the midpoint of the sequence of sorted values. If we observe three numbers: $\{1, 2, 3\}$, the median is 2, even if the numbers are presented as: $\{2, 1, 3\}$. If there is no direct median, we take the average of the two numbers that surround the median. So if we observe: $\{1, 2, 3, 4\}$, the median is 2.5.

Mode

The **mode** of a variable is the value that appears most often in the sequence.

Example

Suppose AT&T is interested in learning about how its customers use their monthly minutes. Suppose AT&T randomly selects 12 customers and observes the following sequence of minutes per month: $\{90, 77, 94, 94, 100, 112, 91, 100, 92, 100, 113, 83\}$

1. What is the mean number of minutes? Is this a population or sample mean?
2. What is the median number of minutes?
3. What is the modal number of minutes?

Measures of location

Note the following properties:

- If mean $=$ median $\Rightarrow$ the distribution is symmetric
- If mean $>$ median $\Rightarrow$ the distribution is right-skewed
- If mean $<$ median $\Rightarrow$ the distribution is left-skewed

Measures of dispersal

The **range** of a variable is the difference between the highest and lowest sorted values.

The **population variance** is:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$$

The **population standard deviation** is:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}$$

where μ is the population mean, and N is the size of the population

Measures of dispersal

The **sample variance** is:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

The **sample standard deviation** is:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

where $\bar{x}$ is the sample mean, and n is the size of the population

Example

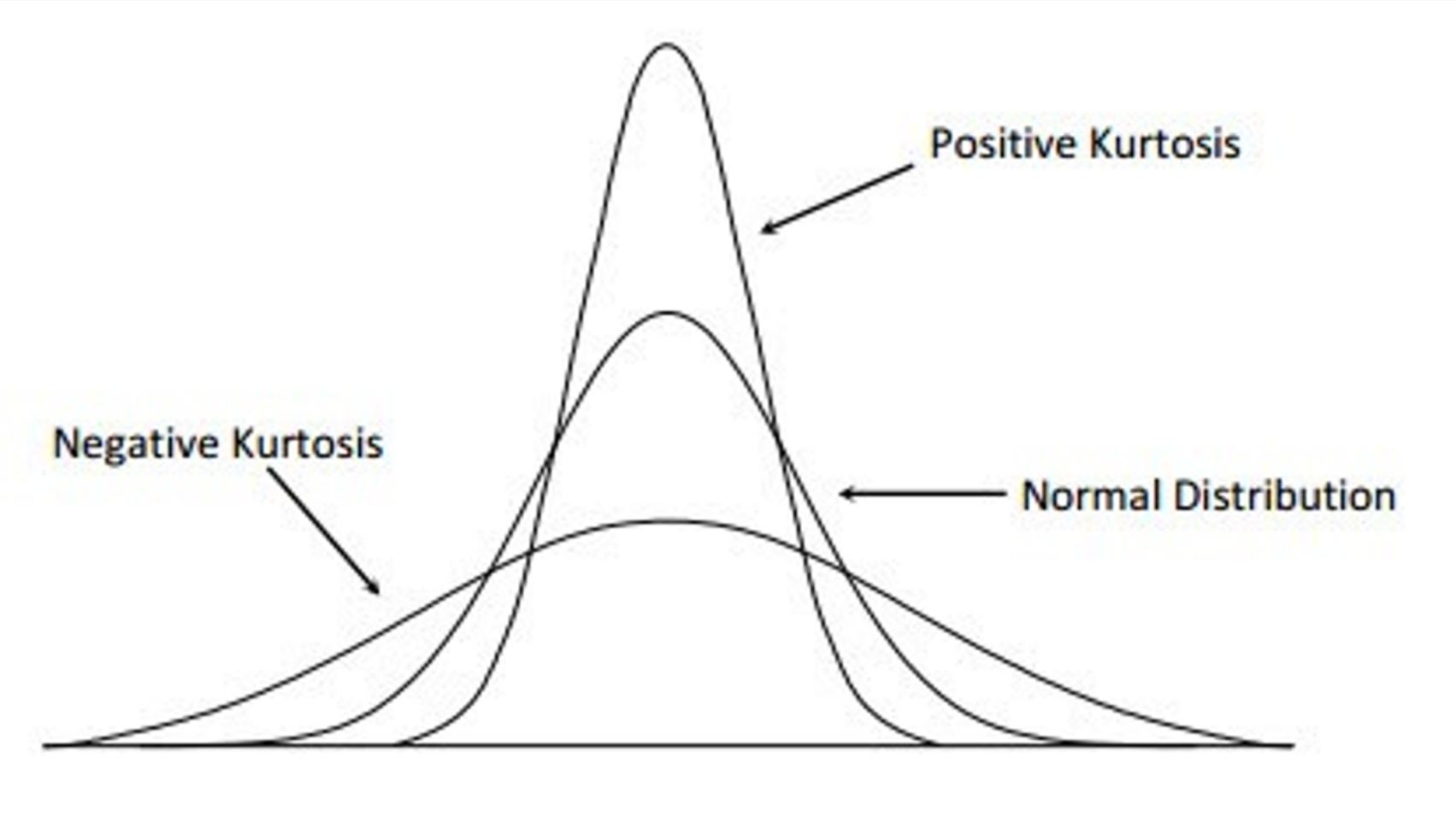
The five vice-presidents of Milton Bradley make the following salaries (in thousands of USD): $\{125, 128, 122, 133, 140\}$.

1. What is the range of these salaries?
2. Calculate the variance and standard deviation of the measures
3. Are these measures a sample or population measure?

Measures of shape

Kurtosis is the measure of the tails of a distribution, compared to the normal distribution. We can calculate it for a variable X by first *standardizing* X into $Z = \frac{X-\mu}{\sigma}$, then calculating: $E\left[\frac{(X-\mu)^4}{\sigma^4}\right]$. The direct calculation is difficult to interpret, so in many applications it may be sufficient to just visually analyze the shape of the distribution.

Kurtosis


Positive Kurtosis
Negative Kurtosis
Normal Distribution

Relationship between variables

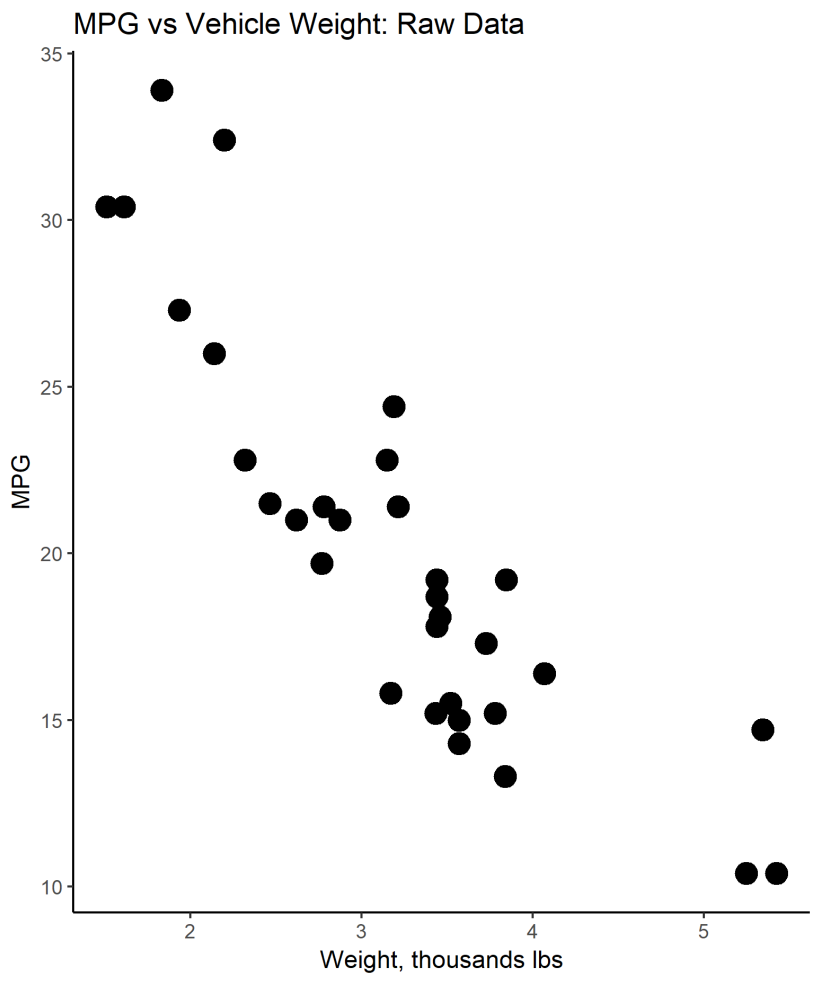
As economists, we are often interested in the relationship between variables. If you work for the Texas Parks and Wildlife Department, perhaps you are interested in the relationship between the price of fishing permits and the stock of fish at a given lake. As a sales manager, you may be interested in the relationship between sales and characteristics of your customers.

A **scatter plot** is a graphical way to represent the relationship between two variables.

The **correlation** and **covariance** between two variables is a numerical measure of the relationship.

Scatterplot

In a dataset with multiple variables, we can plot the relationship between any two variables using a scatterplot, which is nothing more than individual dots representing single observations.



Relationship between variables

If you have two variables, X and Y, then the **correlation coefficient** is a measure of how related the variables are. You can calculate the correlation between X and Y:

$$corr_{X,Y} = \rho_{X,Y} = \frac{1}{(n-1)} \cdot \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{s_X s_Y}$$

Where s_X and s_Y are the sample standard deviations of X and Y, respectively. Note that: $-1 \leq \rho_{X,Y} \leq 1$, and if $\rho_{X,Y} = 0$, there is no correlation between the variables.

Relationship between variables

If you have two variables, X and Y, then the **covariance** is a measure of how related the variables are. You can calculate the covariance between X and Y:

$$cov_{X,Y} = \frac{1}{(n-1)} \cdot \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

Note that while the correlation coefficient is scale-independent (will always be between -1 and 1), covariance depends on the scale of the variables. If $cov_{X,Y} > 0$, then (on average), when X is above its average, Y is above its average. If $cov_{X,Y} < 0$, then when X is above its average, Y is below its average.

Statistical testing

Once we describe data and random variables, we may want to perform statistical tests.

Normal distribution

The **normal distribution** is a common statistical distribution. Check out this great video of a Galton Board that shows the cool properties of the normal distribution. The **probability density function** of the normal distribution is:

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

So the probability of X taking some value x depends on the mean and standard deviation.

Z-test

Using the normal distribution, we can calculate the **z-score** of a single sample mean: how many σ's away from μ.

$$z = \frac{(\bar{x} - \mu)}{\frac{\sigma}{\sqrt{n}}}$$

If $z > 0$, then $\bar{x}$ is above the population mean, and vice versa.

Z-test

We use a Z-test when the sample size (n) is greater than thirty, and we know the population mean and standard deviation.

Example

Suppose the average height of college students is 5ft 9in, with a standard deviation of 7 inches. If we took a random sample from this class, we can test whether our sample mean is different from the population mean.

Z-test

The z-statistic gives us the proportion of the population with values less than the observed sample. A z-score of 0 means the sample and population are identical, so half of the population lies to the left of the sample distribution.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997

t-test

The real world, unfortunately, is messy; we almost never know the population mean and standard deviation. In that case, we must use a **t-test** to compare means. Since we don't know the population parameters (recall the earlier definition of parameters), in the t-test world we are comparing different samples: is the mean of one sample different from another? Using the same example, if we took two samples of height in this class, would the two sample means be equal?

t-test

The t-test (or *Student's t-test*) comes from William Sealy Gosset, the master brewer of Guinness. His mission: determine which variety of barley had the highest yield, to subsequently use to brew Guinness. He observed different fields of barley, and compared the sample means: the world's first t-test!

The t-test

There are several flavors of t-tests:

- One-sample t-test
- Two-sample, equal n, equal variance
- Two-sample, ambiguous n, equal variance
- Two-sample, ambiguous n, unequal variance

The t-test

Two-sample t-test with equal n and equal variance:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s^2}{n}}}$$

Two-sample t-test with ambiguous n and unequal variance:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

The calculated t-statistic falls somewhere on the *t-distribution*: an approximately normal distribution symmetric around zero that changes with the **significance level** and the **degrees of freedom**.

Degrees of freedom

Degrees of freedom are confusing. It is the number of values in the random variable that are allowed to vary while solving for some statistic.

- Two-sample t-test with equal n and equal variance has $2n - 2$ degrees of freedom
- Two-sample t-test with ambiguous n and unequal variance has the following degrees of freedom:

$$d.f. = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}$$

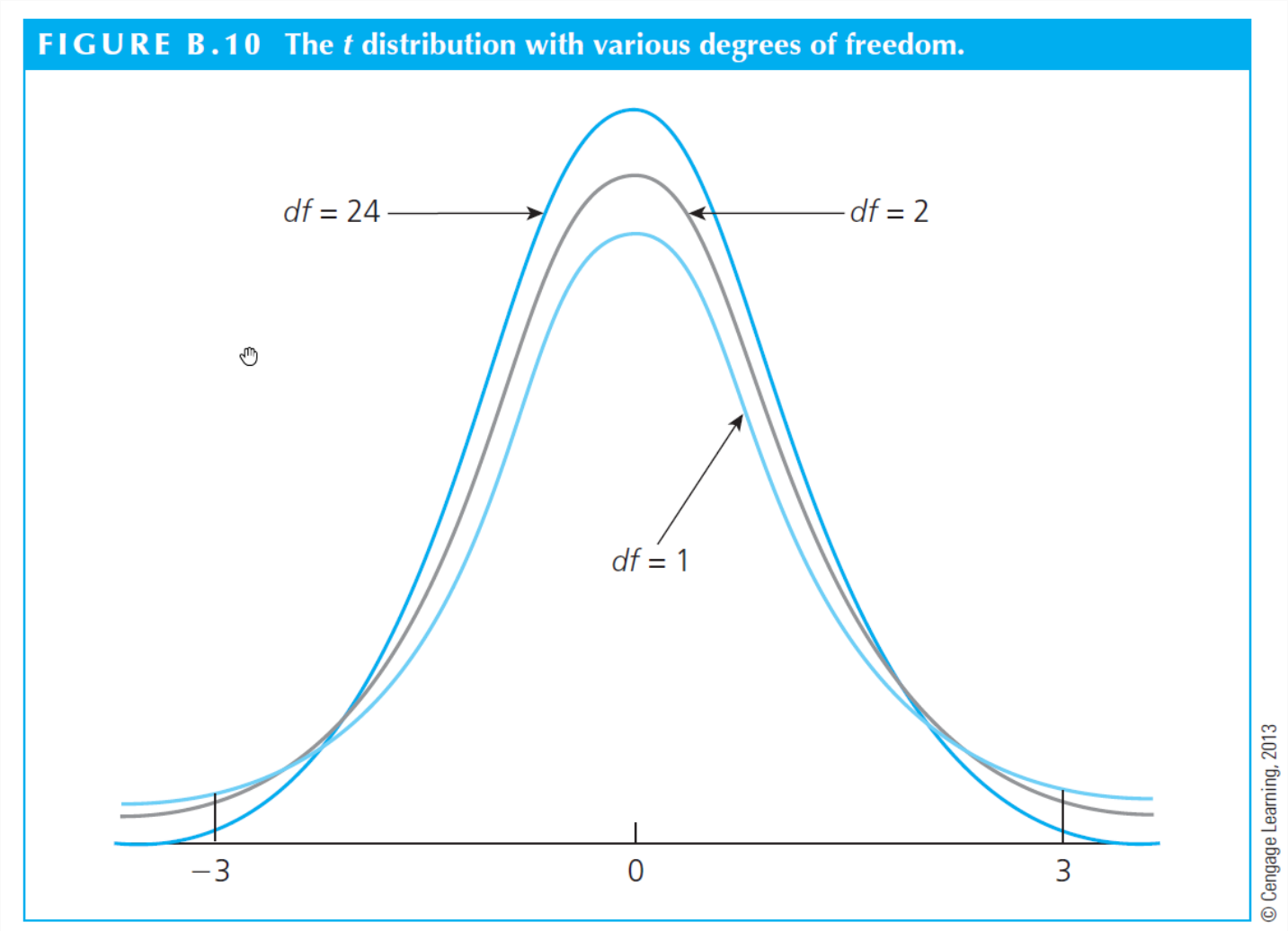
F-test

If we are performing a two-sample t-test, should we assume unequal or equal variances? We can test for that! An **F-test** is a statistical test of the equality of variance across two random independent samples. The F-statistic for two random variables X and Y is:

$$F = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}{\frac{1}{m-1}\sum_{i=1}^{m}(y_i - \bar{y})^2}$$

t-distribution

This is a t-distribution:

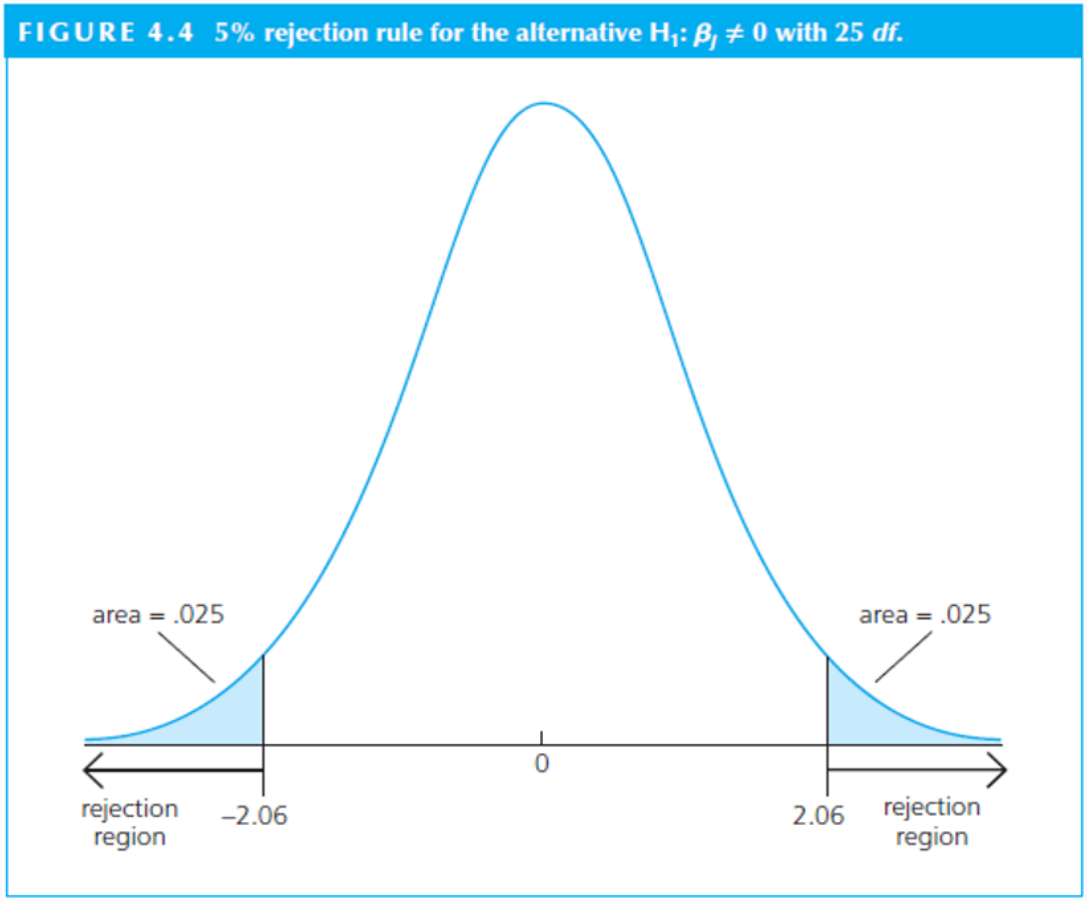


t-distribution

To determine if the difference in means between the samples is significant, we want to see where our t-stat lands on the t-distribution.

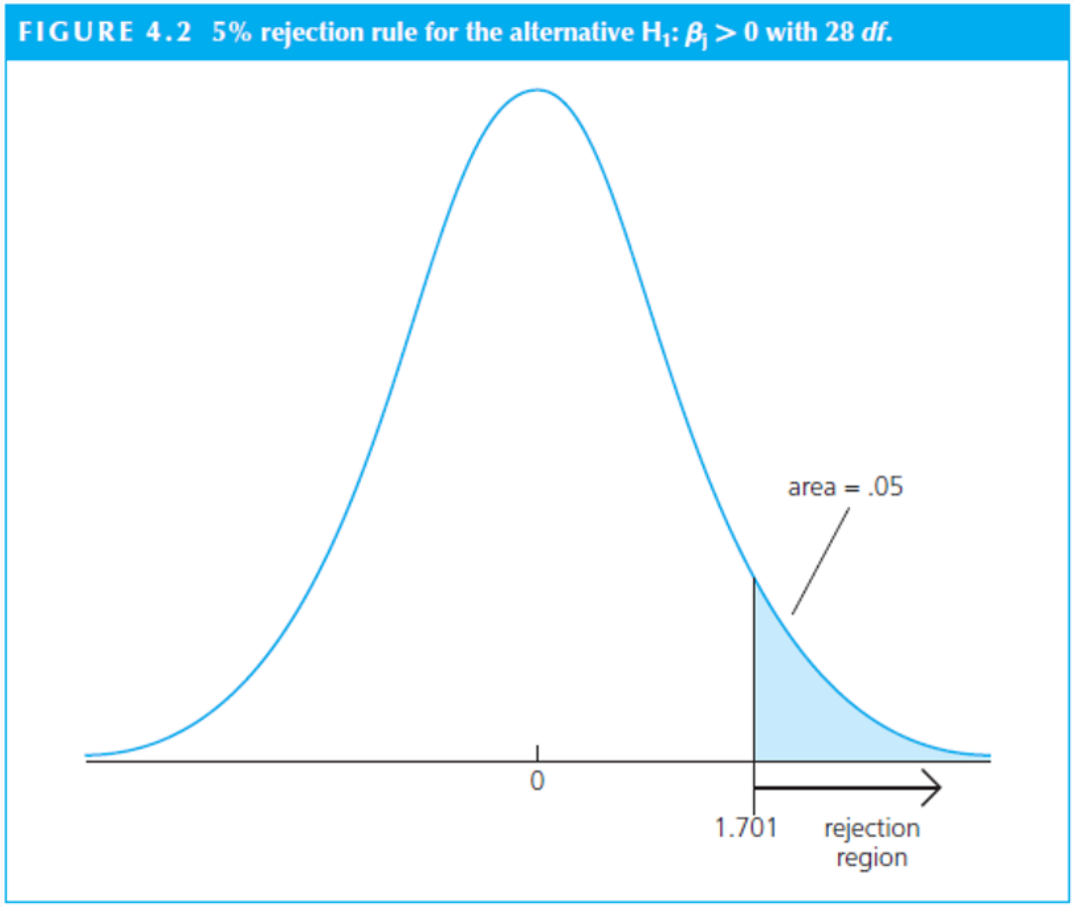
t-test: Visual

Two-tailed test



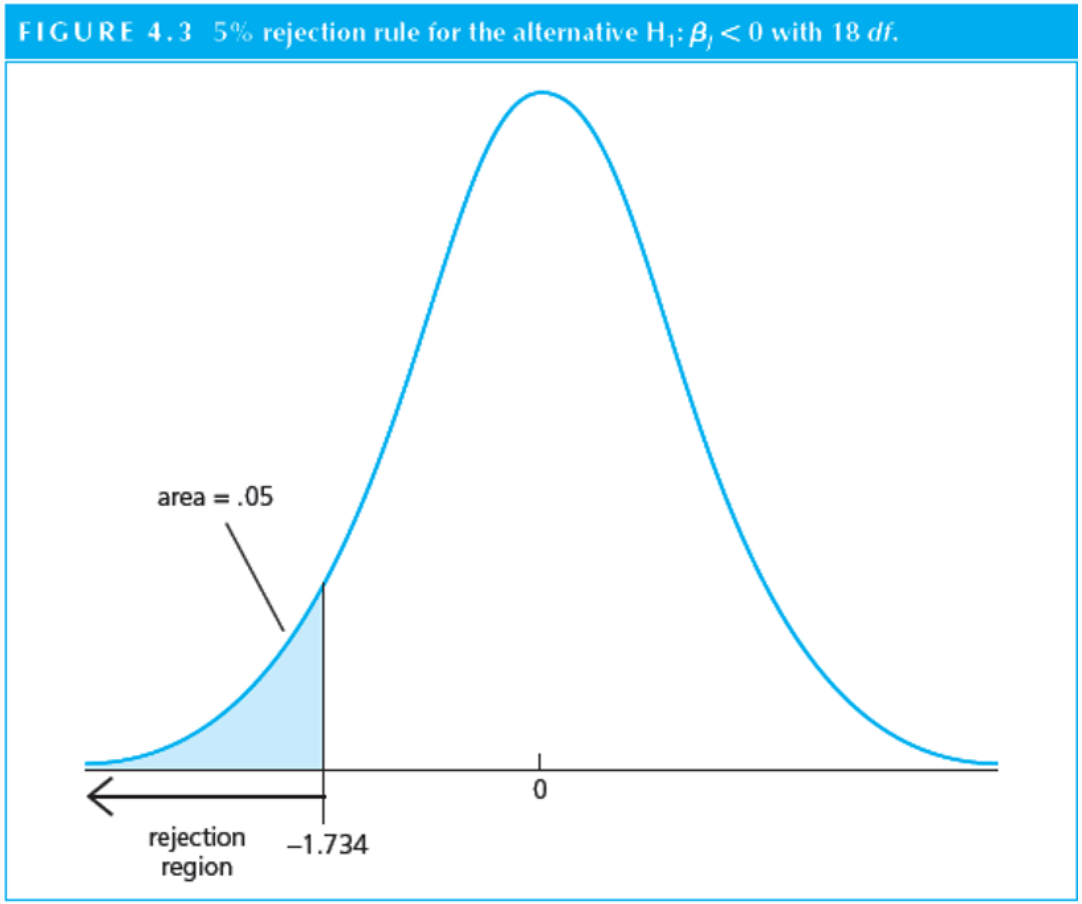
t-test: Visual

One-tailed test: positive



t-test: Visual

One-tailed test: negative



t-test: steps

1. Form your null hypothesis
2. Form the alternative hypothesis
3. Compute t-stat
4. Use α, critical value to make decision to reject or fail-to-reject null.

Statistical hypotheses

We need to specify a **null hypothesis** and an **alternative hypothesis**:

- Null hypothesis (H_0): the default (and boring) guess that nothing is happening, there are no relationships, and **there is no difference between sample means.**
- Alternative hypothesis (H_1): the hypothesis that contradicts the null

$$H_0 : \bar{X}_1 = \bar{X}_2$$
$$H_1 : \bar{X}_1 \neq \bar{X}_2$$

...what type of hypothesis test is this?

Statistical hypotheses

VERY IMPORTANT: with hypothesis testing, you have two options:

- Reject the null hypothesis
- Fail to reject the null hypothesis

You do NOT have the option to say "the null hypothesis is true", nor "the alternative hypothesis is true". The truth is unknowable. The best we can do with statistics is to have enough evidence that we fail to reject a statement.

Statistical hypotheses

When we reject or fail to reject a null hypothesis, we can make two types of errors:

- **Type I error**: rejecting a true H_0
- **Type II error**: failing to reject a false H_0

Decision	H_0 is true	H_0 is false
Fail to reject H_0	Correct decision ($Pr = 1 - \alpha$)	Type II error ($Pr = \beta$)
Reject H_0	Type I error ($Pr = \alpha$)	Correct decision ($Pr = 1 - \beta$)

Statistical hypotheses

We specify α and β.

- α: the probability of committing a Type I error. How comfortable are we with rejecting a true null? Are we willing to make that mistake 10% of the time? 5%? 1%?
- β: the probability of committing a Type II error. We want to be able to correctly reject a false null, so this is often cast as choosing the **power** of the test: How often can we detect a true effect? That is, $1 - \beta$. If we want to be able to correctly identify an effect 90% of the time, $\beta = 0.1$

Statistical hypotheses

The **p-value** of a t-test is the probability of observing a value *more extreme* than the tested value. If we observe a large p-value, then the probability of observing a more extreme value is likely, suggesting that our null hypothesis is correct. If the p-value is low, that suggests our null hypothesis is wrong.

Statistical hypotheses

Generally speaking, when the p-value is less than α, the difference in means is **statistically significant**.

Statistical hypotheses

Our decision on whether to reject or fail-to-reject can come from two related rules:

- Compare the t-statistic to the t-critical value
- Compare the p-value to the α you set

Statistical hypotheses

The t-critical value can come from published tables:

t Table

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence Level										

Statistical hypotheses

...or from Excel:

- Two-tailed: "=TINV(α, df)"
- One-tailed: "=TINV($2 \cdot \alpha$, df)"

Make decision

Two-tailed test:

- If $|t_{stat}| > t_{crit}$: Reject null
- Otherwise: Fail to reject null

Make decision

One-tailed test ($H_1 : \bar{X}_1 - \bar{X}_2 > 0$):

- If $t_{stat} > t_{crit}$: Reject null
- Otherwise: Fail to reject null

One-tailed test ($H_1 : \bar{X}_1 - \bar{X}_2 < 0$):

- If $t_{stat} < -t_{crit}$: Reject null
- Otherwise: Fail to reject null

p-value revisited

The p-value of a coefficient is technically:

Two-tailed: $pvalue = Pr(|t_{stat}| > t_{crit})$

One-tailed, H_1 positive: $pvalue = Pr(t_{stat} > t_{crit})$

One-tailed, H_1 negative: $pvalue = Pr(t_{stat} < -t_{crit})$

We can calculate the p-value in Excel:

"=TDIST(ABS(tstat),df,tails)"

If p-value $< \alpha$, we reject the null hypothesis. Otherwise, we fail to reject the null.

The power of Excel

You can use the information in this section to perform a t-test by hand. Or, you can use Excel. Let's look at an example...

Recap

This lecture should feel dense, but not unfamiliar; much of it should be a review of your statistics classes. Moving forward, you will not be burdened with much new statistical knowledge, but you will be expected to apply basic hypothesis testing to economic situations.

Key skills

After completing this lecture and problem set, you should be comfortable with:

- Summarizing data using tables and figures
- Applying a t-test to test the equality of means for two samples

Going further

If you would like to extend this material, or are serious about becoming an analyst, email me to start using R.