

DataSci 520

lesson 6

hypothesis testing



PROFESSIONAL &
CONTINUING EDUCATION

UNIVERSITY *of* WASHINGTON

today's agenda

- basic hypothesis tests
- types of hypothesis tests
- steps in a test of hypothesis
- multiple hypotheses and P-value fishing

Statistical Hypotheses

- > A statistical hypothesis is a statement that *can be directly measured or tested*.
- > A full and complete hypothesis statement covers ALL outcomes of measured events.
- > Every statistical hypothesis test is stated completely BEFORE outcomes are measured.



Excellent health statistics - smokers are less likely to die of age related illnesses.'

Hypothesis Testing

- > Identify a hypothesis that can be tested.
 - “Changing our web-site logo to be bigger on the front page will drive more than 100,000 customers to our site per day.”
- > Select a criteria to evaluate the hypothesis.
 - If our sample has a probability of $\geq 90\%$ chance that there are more than 100,000 customers per day, we accept the hypothesis.
- > Select a random sample from the population.
 - Randomly assign a cookie to new site users that tells the server to show A or B website.
- > Compare observations to what we expect to observe and calculate statistic and the resulting probability.

Hypothesis Testing

- > We first state our population assumptions in the null hypothesis. ($H_0 = "H - Not"$)
- > We state our new alternative hypothesis as an alternative to the null. ($H_a = "H - a"$)
 - The null + alternative should make up all possible outcomes and be mutually exclusive

H_0 : "The old website drives equal amount of traffic or more."

H_a : "The old website drives less traffic than the new one."

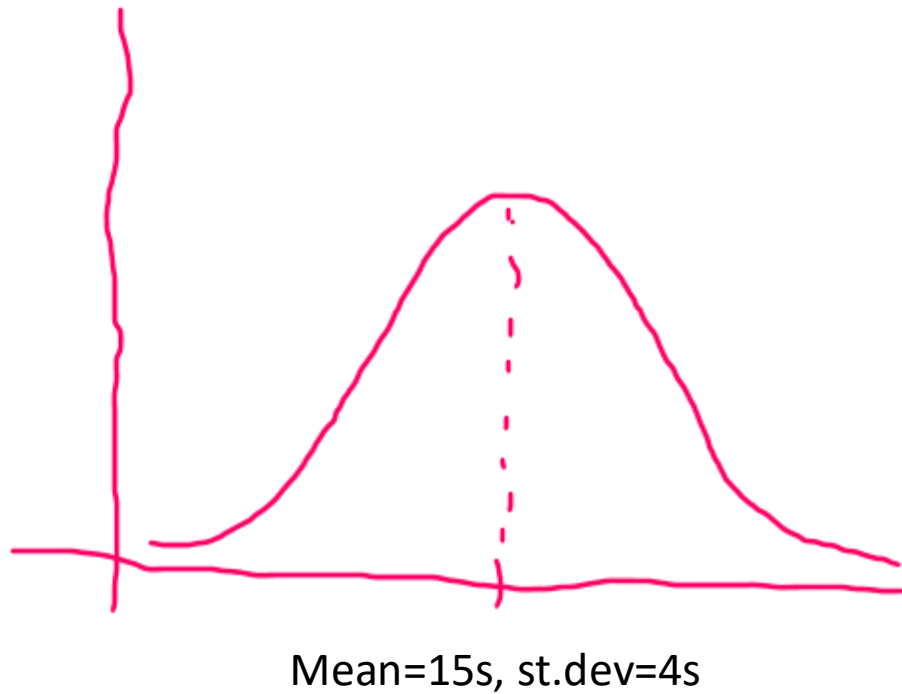
- > Decide on a significance level (probability cutoff)
 - 0.9, 0.95, and 0.99 are common (problem specific)

Hypothesis Testing

- > Based on our findings we can only do two things
 - We reject the null-hypothesis.
 - > Since the alternative covers all other possibilities, we can say we accept the alternative hypothesis.
 - We fail to reject the null hypothesis.
 - > We do not accept the null hypothesis because it is really really hard to prove something is absolutely true. Instead maybe we just got unlucky.
 - > We could have failed for other reasons:
 - The alternative hypothesis was false to begin with.
 - We did not collect enough evidence for the alternative hypothesis.

Hypothesis Testing - Example

- > We know that the average time a user spends on a page has a mean of 15 seconds and a st.dev. of 4 seconds.
 - If we assume normality, how do we test if a change to the page has a higher view time?



Formulating the Hypothesis Statement

- > We know that the average time a user spends on a page has a mean of 15 seconds and a s.d. of 4 seconds.
 - If we assume normality, how do we test if a change to the page has a higher view time?

H_0 : The old website has the same or more viewership than the new website.

H_a : The old website has less viewership than the new website.

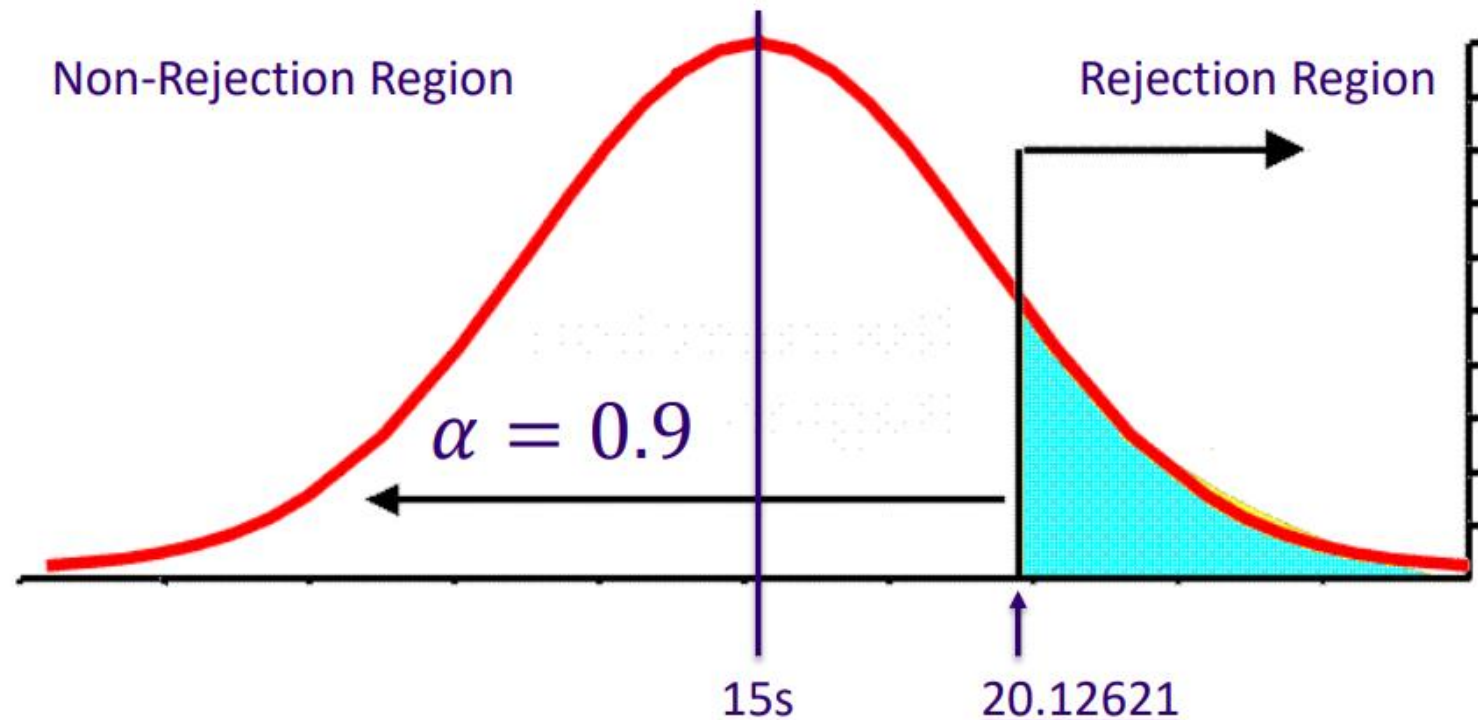
OR (the same one)

H_0 : The new website has the same or less viewership than the old website.

H_a : The new website has more viewership than the old website.

Hypothesis Testing

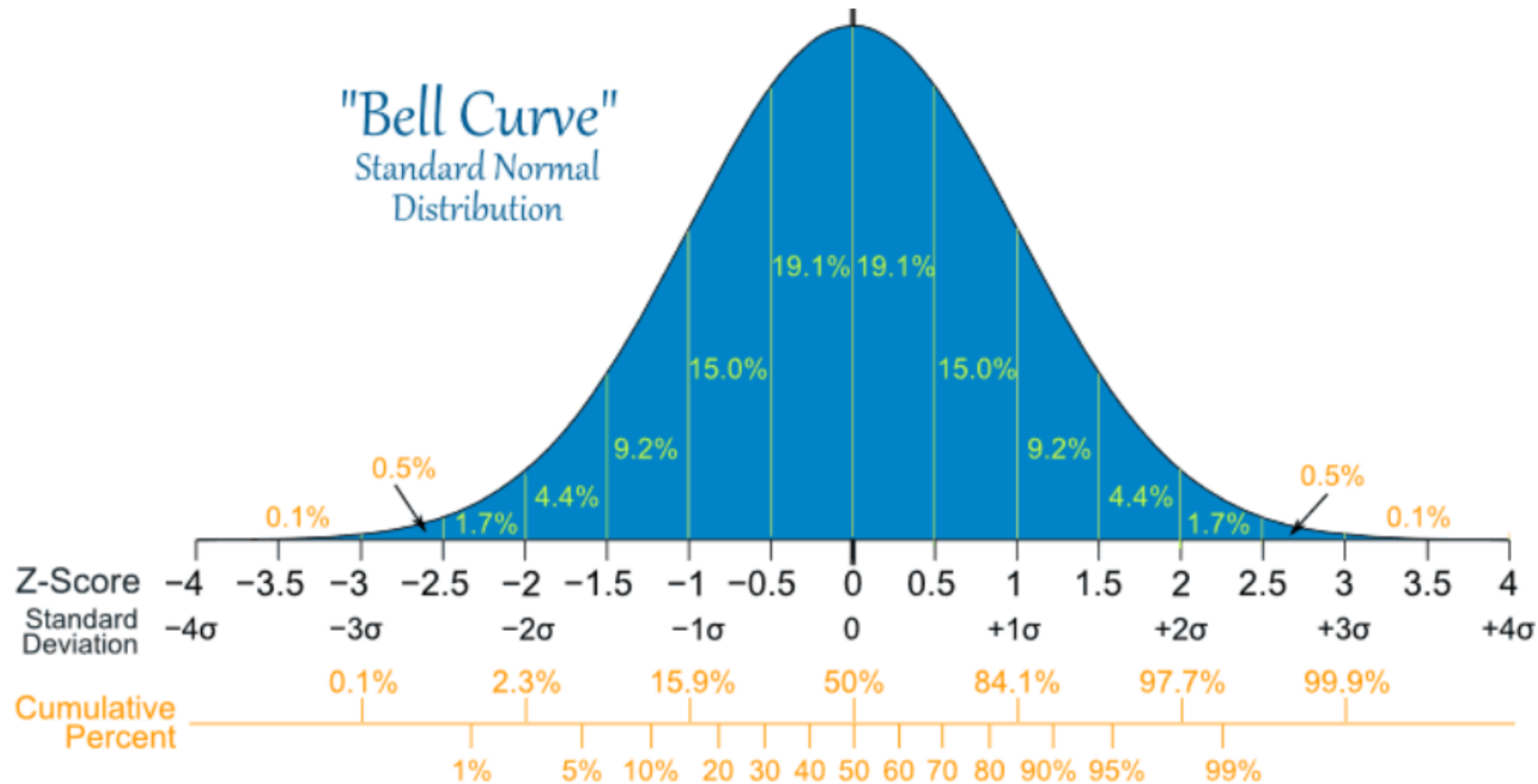
- > We now select a confidence value.
 - For example, we select 90% confidence.
 - Then an event in the blue region will have a 10% chance or less of occurring.



from scipy.stats import norm
norm.ppf(.9, loc=15, scale=4)

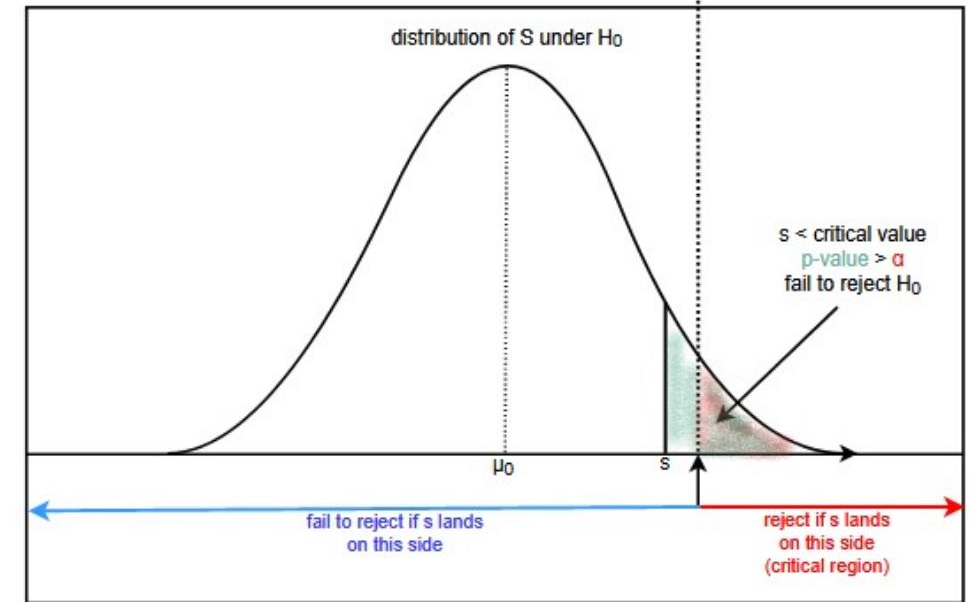
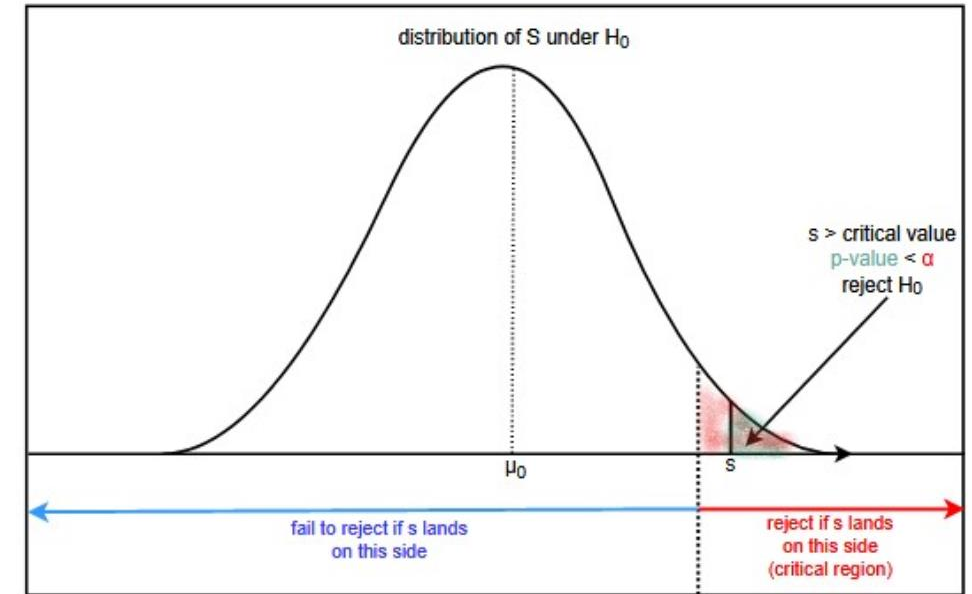
FYI – Normal Distribution (Gaussian Distribution)

- > Areas (probabilities) on the normal distribution are directly related to the distance from the mean.



one-tailed test of hypothesis

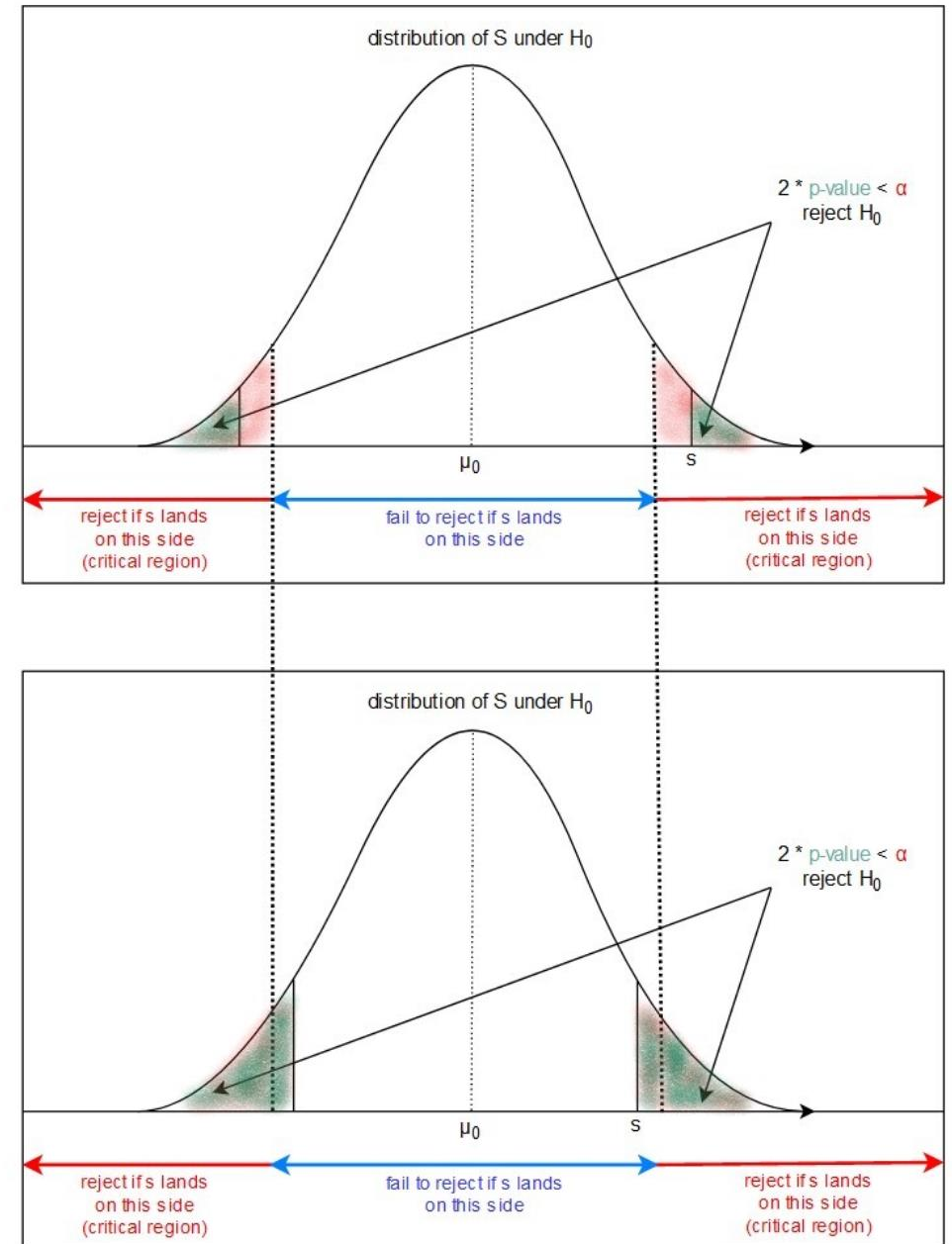
- $H_0 : \mu = \mu_0$ vs $H_a : \mu > \mu_0$
- if the statistic falls in the **critical (red) region**, we reject H_0 because $p\text{-value} < \alpha$ (top graph)
- if the statistic falls in the **blue region**, we fail to reject H_0 because $p\text{-value} > \alpha$ (bottom graph)



one-tailed

two-tailed test of hypothesis

- $H_0 : \mu = \mu_0$ vs $H_a : \mu \neq \mu_0$
- if the statistic falls in the **critical (red) region**, we reject H_0 because $2 \times \text{p-value} < \alpha$ (top graph)
- if the statistic falls in the **blue region**, we fail to reject H_0 because $2 \times \text{p-value} > \alpha$ (bottom graph)



steps involved in a test of hypothesis

1. State the null hypothesis H_0 , the alternative hypothesis H_1 and choose α
2. Compute the statistic. How to compute the statistic depends on the test. Every test has a different calculation.
3. Compute the p-value. For a two-tailed test, we also multiply by 2.
4. If the p-value is less than α , then we reject H_0 . Otherwise we fail to reject H_0 .

Doing steps 2 and 3 by hand is usually tedious and error prone, so in Python we can look up which function to call in `statsmodels` or similar packages in order to perform a given test. We can pass the data directly to such functions and they will do all the calculations and report the p-value to us.

Hypothesis Testing – Defining the P-Value

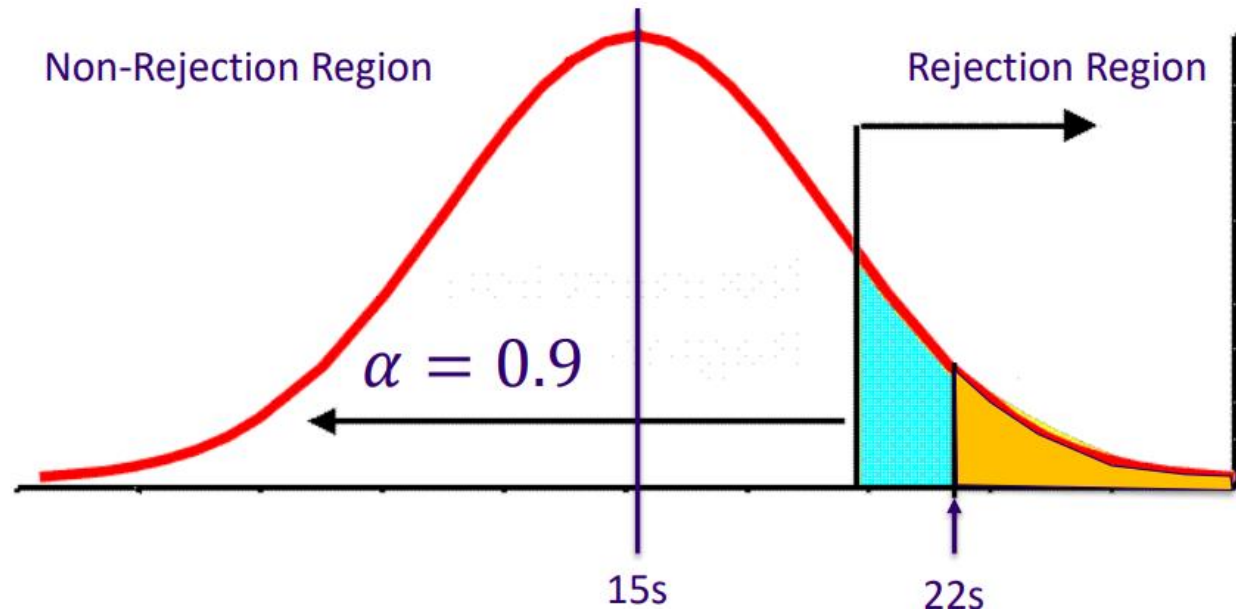
- > To test the hypothesis we created, we perform an experiment.
- > The experiment usually results in creating a **test-statistic** from **a sample**.
 - E.g. “The new website resulted in **5,580 users** viewing the page for an **avg of 22 seconds**.”
- > We want to answer the question:

“What is the probability of observing our test statistic under the assumption that the null-hypothesis is true?”

- > We then arrive at that exact probability... which is called the “P-value”.
- > If the p-value is small, then we can say:
 - This sample has too low of a probability of occurring by random chance.
 - This means there is enough statistical evidence to accept the alternative hypothesis.
- > If the p-value is too large, then we can say:
 - This sample may have happened by chance.
 - There is not enough statistical evidence to accept the alternative hypothesis.

Hypothesis Testing – P-Value

- > The p-value is the probability of obtaining the sample results or worse.
- > What is the p-value of a sample mean of 22 seconds?

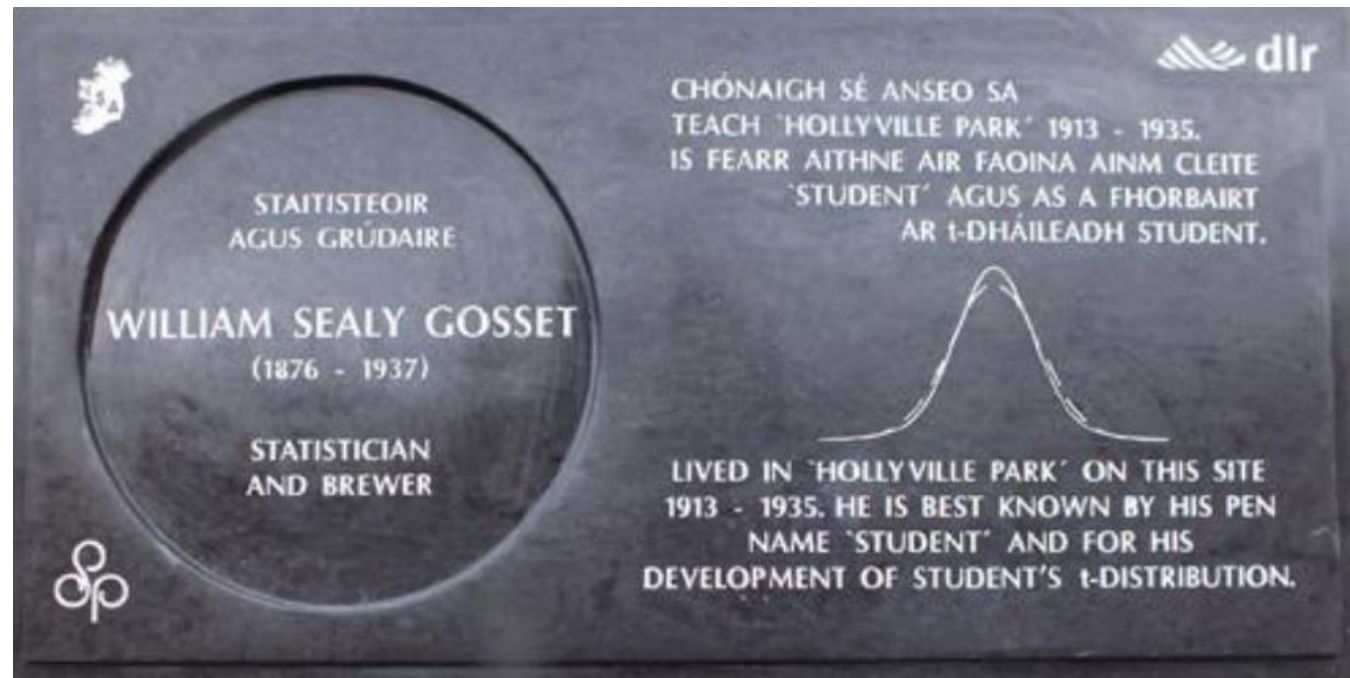


Cutoff Region
P-value Probability

```
from scipy.stats import norm  
1 - norm.cdf(22, loc=15, scale=4)  
>> 0.04006=~4%
```


More Hypothesis Testing: What if we don't know our pop's St.Dev?

- > Sometimes we do not know our population's standard deviation.
 - The population is hard to observe: (Remember, not everything is just a log-statement away)
 - > E.g. "The amount of parasites in the GI tract of an endangered animal population."
- > To address this, we can think of using something like the normal distribution, but with longer tails to account for the possibilities of outliers. (Heavier Tails)
- > The distribution has a very in-depth statistical derivation... BUT
- > The result is the "Student's T Distribution".



discussion

Many websites list different tests of hypothesis and when to use them. Here are two examples:

- [List of hypothesis tests \(UCLA\)](#)
- [List of hypothesis tests \(machinelearningmastery.com\)](#)

Take a moment to visit one of the above sites and look at the different tests.

- What kind of parameter is being tested in each case?
- Can you spot the null and alternative hypotheses?

Hypothesis Testing: T-Test

- > Student's T-test: tests a hypothesis about the difference of two sets of data:
 - Test whether a population mean has a specified value.
 - Test the difference between two means (equal, unknown variances).
 - Test a paired-response difference from zero.
 - > E.g. a before/after drug treatment on patients.
 - Test whether the slope of a line is not zero.
 - > Important for testing the importance of variables (later in class).
- > Use 'Welch's T-test' for testing the difference between two means (unknown variances, potentially different).
- > Picking the different tests changes test's results.
- > The more assumptions we make, the easier it is to tell the difference between populations.

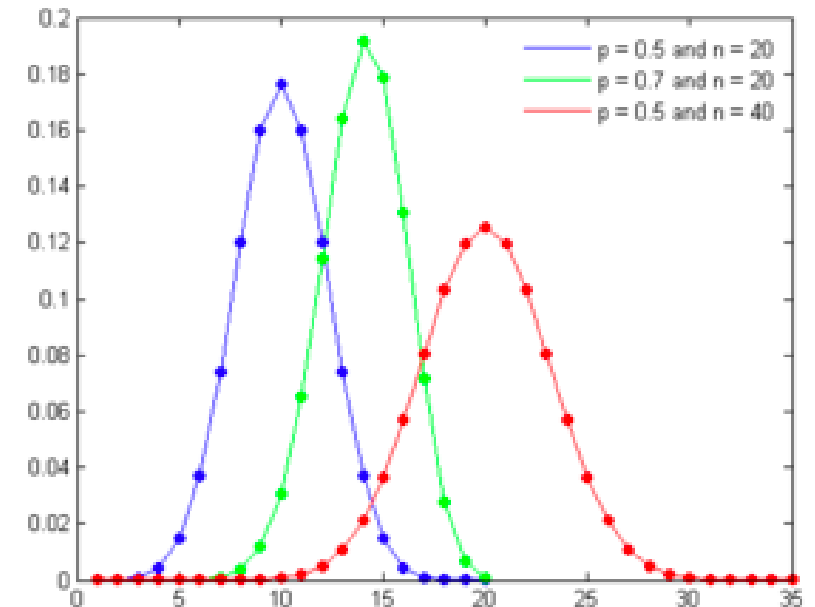
Hypothesis Testing: What about discrete data? (Counts...)

- > The prior tests assume the possibility of observations coming from a real-number line (or subset of it).
- > What about counts or observations of discrete integer outcomes?
- > Yes, we can do this...
 - Imagine flipping a coin 40 times. We can ask: Is the coin fair?
 - In other words:

H_0 : The coin is fair. The count of “heads” is half.

H_a : The coin is not fair. The count of “heads” is not half.

So the Null Hypothesis distribution for $N=40$, $p=0.5$ is →



Hypothesis Testing: What about discrete data? (Counts...)

H_0 : The coin is fair. The count of "heads" is half.

H_a : The coin is not fair. The count of "heads is not half.

Step 1: Pick a confidence level. Let's pick 0.95 (95%).

- If we observe something that has $\leq 5\%$ chance of occurring, we will accept alternative hypothesis (coin is not fair).

Step 2: Perform experiment. Let's say we observe 14 heads (and 26 tails). Is there enough evidence to reject the claim, "The coin is fair"?

Step 3: Breakout Discussion!

Paired Outcomes

- > Sometimes count data can be Paired, meaning we are counting the same object or individual before **and** after an event. Examples
 - Migraine treatment: # Migraines before vs # Migraines after treatment.
 - User Online Social: # of Likes before Ad vs # of Likes after Ad.
 - Data Science Class Members: # of students in class before hypothesis testing vs after.
- > The idea here is that the NULL hypothesis states some form of “no effect”.
 - If no effect, then the distributions (both before and after) should come from the same underlying population distribution.
- > We will compute a statistic that tells us just how far off each group is from before and after.

Categorical Groups! Counting Successes in different groups.

- > Use Fisher's Exact test...
- > E.g. Let's say we have 2 groups (Patients in treatment A vs B). The outcome is patients relapsing in a symptom.

	Cat. A	Cat. B
Success	w	x
Failure	y	z

- > Hypothesis: (Assume we know the total # of successes)
- > Null: The proportion of successes in Category A is not less than that in Category B.
- > Alternative: The proportion of successes in Category A is less than in Category B.
 - We specifically observe the following outcome:
- > We observe 12 patients:
 - 5 in Category A and 7 in Category B.
 - 5 total Successes and 7 total Failures.
 - Here are all the potential ways that can happen:

	Cat. A	Cat. B
Success	2	3
Failure	3	4

0	5	1	4	2	3	3	2	4	1	5	0
5	2	4	3	3	4	2	5	1	6	0	7

Categorical Groups! Fisher's Exact test

- > Hypothesis: (Assume we know the total # of successes)
- > Null: The proportion of successes in Category A is not less than that in Category B.
- > Alternative: The proportion of successes in Category A is less than in Category B.
- > We observe 12 patients:
 - 5 in Category A and 7 in Category B.
 - 5 total Successes and 7 total Failures.
 - Here are all the potential ways that can happen:

	Cat. A	Cat. B
Success	2	3
Failure	3	4

This is the exact outcome we observed.

0	5	1	4	2	3	3	2	4	1	5	0
5	2	4	3	3	4	2	5	1	6	0	7


This is how many ways we can have our observation OR “more extreme” under the alternative hypothesis.

Categorical Groups! Fisher's Exact test

- > Hypothesis: (Assume we know the total # of successes)
- > Null: The proportion of successes in Category A is not less than that in Category B.
- > Alternative: The proportion of successes in Category A is less than in Category B.
- > Observation:

	Cat. A	Cat. B
Success	2	3
Failure	3	4
- > What are the probabilities of each of these observations?

	Cat. A	Cat. B
Success	2	3
Failure	3	4



These column counts are fixed or assigned.

So the probability that this exact distribution occurs under these column counts is:
(5 choose 2 successes) * (7 choose 3 successes) / (12 choose 5 successes)

$$= \frac{\binom{5}{2} * \binom{7}{3}}{\binom{12}{5}} = \frac{10 * 35}{792} = 0.4419$$

Categorical Groups! Fisher's Exact test

> Probabilities for each outcome:

0	5	1	4	2	3	3	2	4	1	5	0
5	2	4	3	3	4	2	5	1	6	0	7
0.0265		0.2210		0.4419		0.2652		0.0442		0.0013	
Our outcome or more extreme											

> $0.0265 + 0.2210 + 0.4419 = 0.6894$

> In other words: "There is a 68.9% chance of observing our success/failure outcome or more extreme by mere chance."

– This is quite large, and means that we do NOT have enough evidence to reject the null.

Chi-Squared Test (Pearson's)

- > Unpaired test for counts in different categories.
- > These categories must be mutually exclusive.
 - Does the patient have cancer? (yes/no)
 - > Test if the two categories differ in WBC count.
 - Rolling a die. (1,2,3,4,5,6)
 - > Test if the six categories occur the same (fair die).
 - Does a tweet contain a specific word? (yes/no)
 - > Test if the two categories differ in tweet length or word count.
- > This tests whether the different categories differ in some specific value.
 - In order to do this test, we have to specify the 'degrees of freedom' in the Chi-squared test.
 - This is equal to $n-1$. Where n equals the number of different categories.
- > The test looks at the sum of the outcome differences from expectations.

Problem Statement

- > Your Aunt has a online shop where she sells hats for pets.
- > She wants to make a change to the checkout process.
- > She (amazingly) keeps great user records of what historically has happened:
 - 60% of customers landing in the cart/checkout page drop off with no purchase.
 - 30% of customers buy.
 - 10% of customers return to shop to continue shopping.
 - This is all known from prior history.
- > She now wants to run an experiment and see if anything changes if
 - She adds a chatbot to pop up and prompt for questions.
- > She will run this experiment for 10 days or until 120 customers land on the checkout page.
- > The final results are in...

Chi-Squared Test (Pearson's)

Given

> Example: A-B test with 3 different outcomes.

	New Occurrence/ Counts	Expectation %	Expectation Counts	Difference	Squared Difference	(Squared Difference) / Expected
Leave Page	55	0.6	$=0.6*120=72$	$=55-72=-17$	289	$=289/72=4.104$
Continue Purchase	43	0.3	$=0.3*120=36$	$=43-36=7$	49	$=49/36=1.361$
Add More to Purchase	22	0.1	$=0.1*120=12$	$=22-12=10$	100	$=100/12=8.333$
Totals	120					13.708

- > If we calculate this type of statistic for each category → then the resulting number comes from a known distribution called the “Chi-Squared Distribution”.
- > Test statistic is 13.708 on a chi-squared distribution with $(3-1)=2$ degrees of freedom.
 - Degree of freedom is # of options minus 1.
 - “`scipy.stats.chisquare([55, 43, 22], [72, 36, 12])`”=13.708, 0.0011
 - Result= Evidence suggests we should reject the null.

Chi Squared Test (Pearson's)

- > Chi Squared Test is also used as a “Goodness of Fit test”. (unpaired)
- > Test if your sample is representative of the expected population.
 - Test if your sample has expected make up of categories.
 - E.g. If we expect that our population is 50-50 men-women, then we test if our sample is different from those expected probabilities (0.5, 0.5).
 - E.g. if we sample individuals/users from all 50 states, we can test if each group is $\sim 1/50^{\text{th}}$ of total sample.
- > BUT if our total sample size is small, we see a breakdown of the Chi squared test. (a subgroup size $\sim < 10$)
 - Then we should switch to a more exact probability calculation (similar to how we did with Fisher's exact test).

Paired Discrete Data (McNemar's Test)

- > Consider the Sickle Cell Anemia Test problem concerning conditional probabilities.
 - Remember that a positive first test could just mean insurance will approve a more accurate & expensive second test.

	# Positive 2 nd Test	# Negative 2 nd Test	Totals
# Positive 1 st Test	80	55	135
# Negative 1 st Test	15	100	115
Totals	95	155	250

- > Usually want to test the “Importance of Having a Second Test”.
- > Null: The probability of negative/positive on 1st test == probability on 2nd test.
- > Alternative: They are not equal
- > OR
- > Null: $p(- \text{ and } 1^{\text{st}} \text{ test}) == p(- \text{ and } 2^{\text{nd}} \text{ test})$
- > Alternative: $p(- \text{ and } 1^{\text{st}} \text{ test}) \neq p(- \text{ and } 2^{\text{nd}} \text{ test})$

Paired Discrete Data (McNemar's Test)

> Data:

	# Positive 2 nd Test	# Negative 2 nd Test	Totals
# Positive 1 st Test	80	55	135
# Negative 1 st Test	15	100	115
Totals	95	155	250

> McNemar's Statistic is:

$$\chi^2(f) \sim \frac{(\#neg_{2nd} - \#neg_{1st})^2}{\#neg_{2nd} + \#neg_{1st}}, \text{ where } f = \text{degrees of freedom} = \# \text{ categories} - 1 = 1$$

$$\chi^2(f = 1) \sim \frac{(55 - 15)^2}{55 + 15} = \frac{1600}{70} = 22.86$$

Scipy.stats.chi2(22.86, df=1)

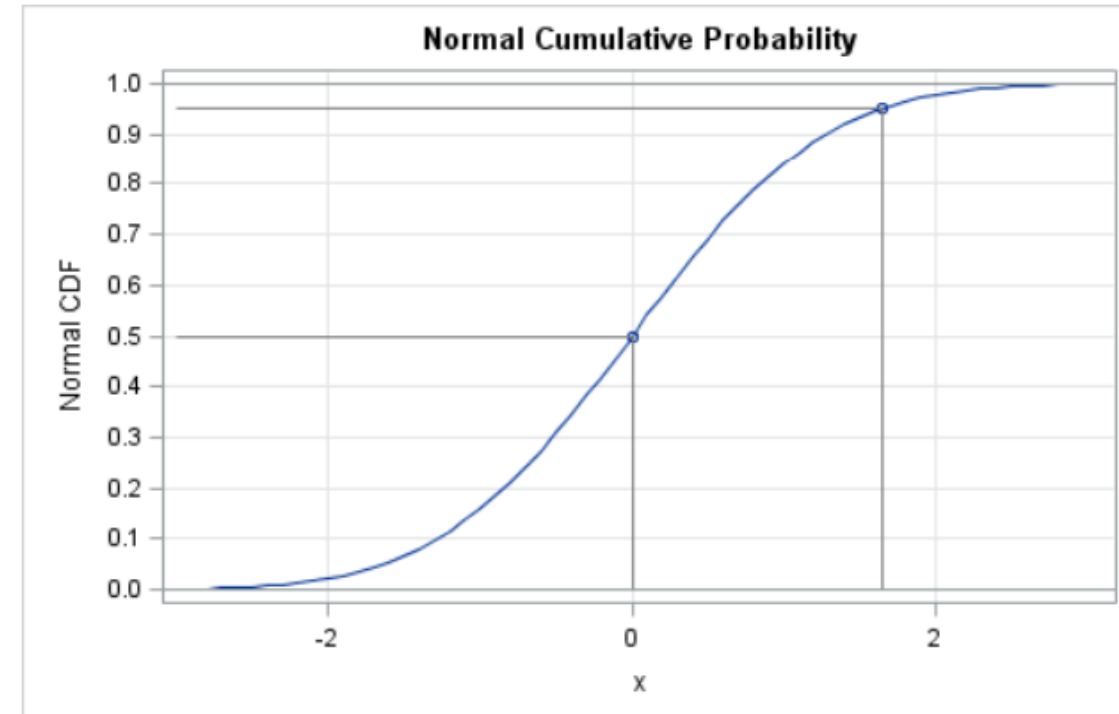
>>9.07e-07

Testing for Normality (IMPORTANT)

- > It is very important to be able to test a feature or set of numbers STATISTICALLY if they have a high probability of coming from a normal distribution.
- > Need to be able to say more than, "The histogram looks normal..." → Not enough. Especially with smaller sample sizes or missing data.

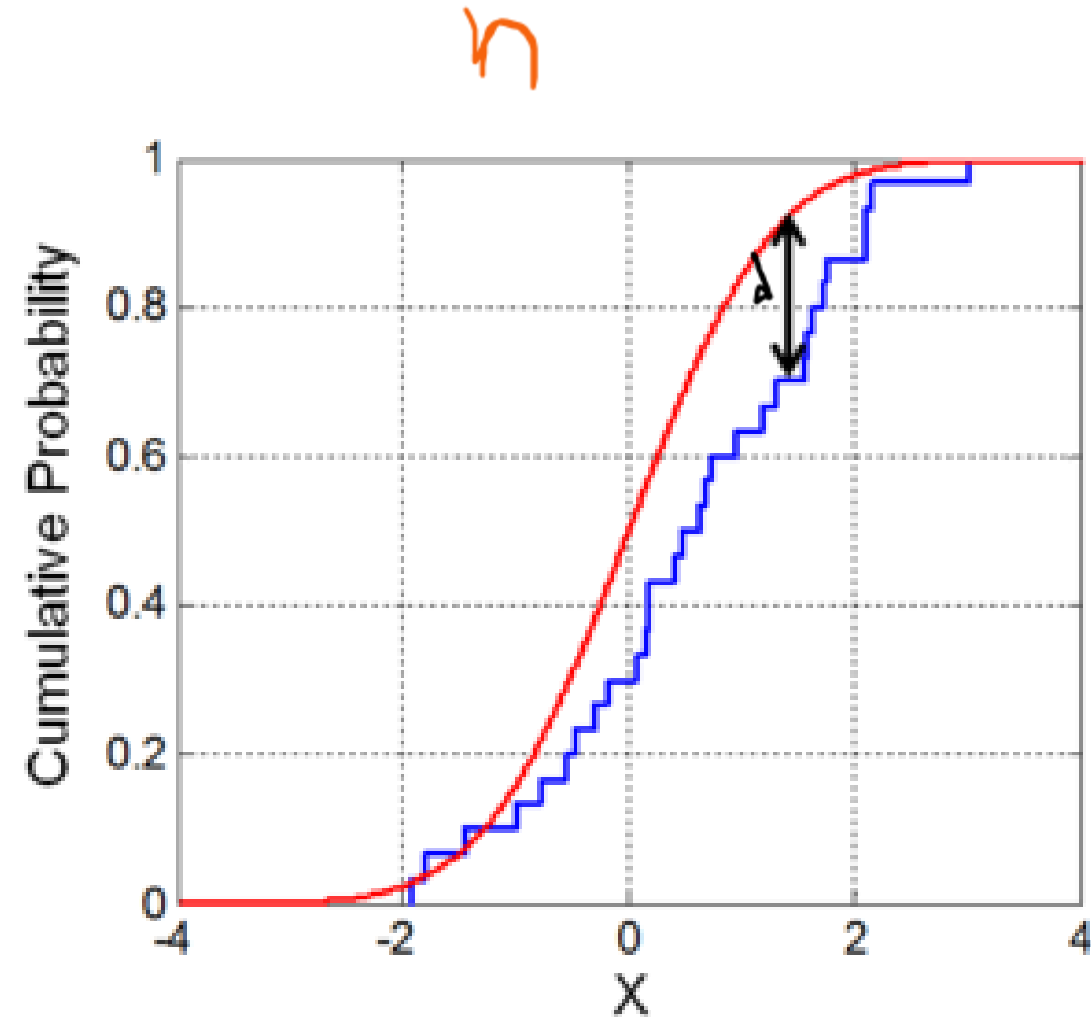
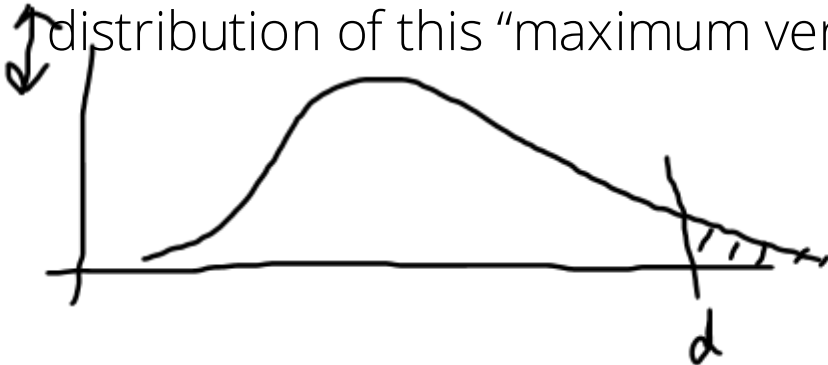
Kolmogorov-Smirnov test (K-S test).

- > Consider the Normal Cumulative Distribution Function (CDF), the sum of the area to the left of x .
- > The plot to the right is the CDF of a standard normal distribution.
- > Idea: Compare this to your sample CDF.
- > K-S Statistic:
 - Maximum vertical distance between the sample and the population CDF.
- > NOTE: this actually works with any population distribution



K-S Normality Test Example

- > The larger the maximum vertical distance, the more likely the **sample** does not come from the general **population distribution**.
- > Problem: How does this computed statistic tell us the probability of normality?
 - It doesn't really.
 - Note: there are some specific edge cases of situations in which statisticians can derive the expected statistic distribution.
 - Instead, we can just do something like bootstrapping! Bootstrapping (sampling w/ replacement) can get us a handle on the distribution of this "maximum vertical distance".

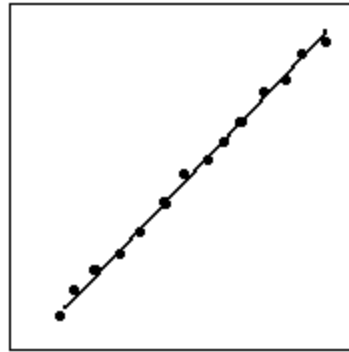


Testing for Normality: Shapiro-Wilk Test

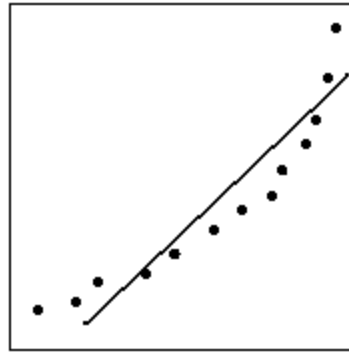
- > The Shapiro-Wilk test can tell us a test statistic for normality.
 - Tests the difference in expected and sample 'moments'.
 - Moments:
 - > 1st moment = mean
 - > 2nd moment = variance
 - > 3rd moment = skewness
 - > 4th moment = kurtosis
 - > ...
 - Slightly more powerful than the K-S test, which means there are more assumptions made (see all the above moments).
 - See: <https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.shapiro.html>

Testing for Normality

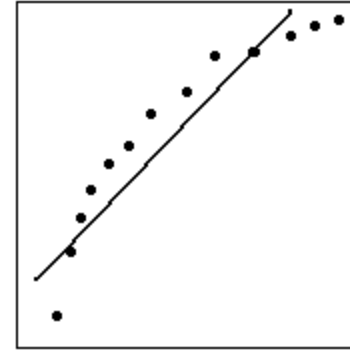
- > Always look at the qq-plot.
 - Plot of the sample quantiles VS hypothesized normal quantiles.



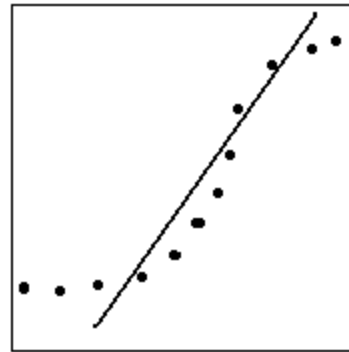
a. Normal



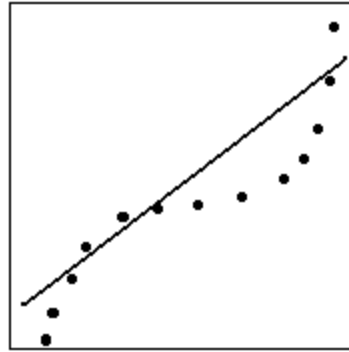
b. Skewed to the Left



c. Skewed to the Right



d. Thick Tails



e. Thin Tails

Hypothesis Testing between Multiple Groups!

- > What if we had multiple groups and we wanted to compare their means?
 - Why can't we just do multiple two-sample t-tests for all pairs?
 - Results in increased probability of accepting a false hypothesis.
 - E.g., if we had 7 groups, there would be $(7 \text{ Choose } 2)=21$ pairs to test. If our alpha cutoff is 5%, then we are likely to accept about 1 false hypothesis ($21*0.05$).

Weekday	Commute Times (minutes)
Sunday	10, 14, 7, 9, 9, 11, 10, 9, 10
Monday	12, 15, 18, 11, 11, 9, 10, 8, 14
Tuesday	11, 13, 13, 9, 8, 11, 12
Wednesday	11, 10, 14, 11, 9, 10, 15, 15
Thursday	15, 17, 19, 10, 14, 14, 9, 10
Friday	19, 14, 14, 10, 17, 9, 12, 13
Saturday	8, 8, 10, 14, 11, 11, 9

21 T-tests to perform:

- Sunday vs Monday
- Sunday vs Tuesday
- Sunday vs Wednesday
- ...
- Wednesday vs Friday
- ...
- Friday vs Saturday

5% cutoff for each. What is the probability of at least one false null rejection? I.e., $1 - \text{probability of all right}$.

$1 - (0.95^{21}) = 0.659... = 66\%$ chance of a false null rejection!!!!

Hypothesis Testing between Multiple Groups!

- > Null Hypothesis:
 - All groups are just samples from the same population.
- > Alternative Hypothesis:
 - At least one group has a statistically different mean.
- > This type of analysis is called “ANalysis Of VAriance”, or ANOVA.
 - We make data independence and normality assumptions first.

- > Our test statistic is based on:

$$statistic \sim \frac{\text{between group variability}}{\text{within group variability}} = \frac{SS_{\text{between}}/(\#groups - 1)}{SS_{\text{Total}}/(\#data - \#groups)} \sim F(\#g - 1, \#d - \#g)$$

- > This is an “F-statistic” and is distributed as a known distribution called an “F-distribution”.
- > The probability of this ratio is our p-value.

****IMPORTANT NOTE**:**

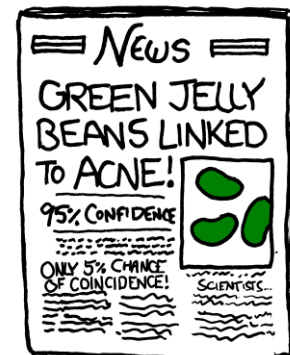
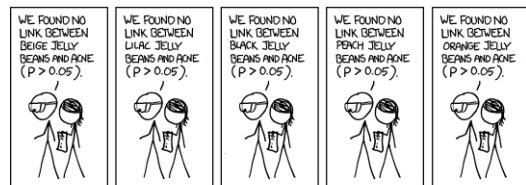
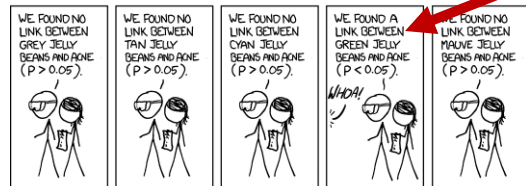
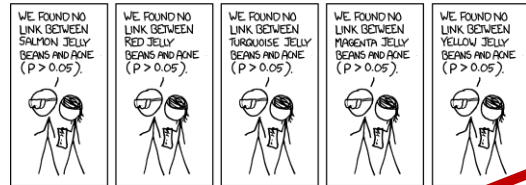
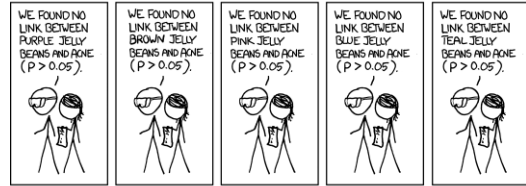
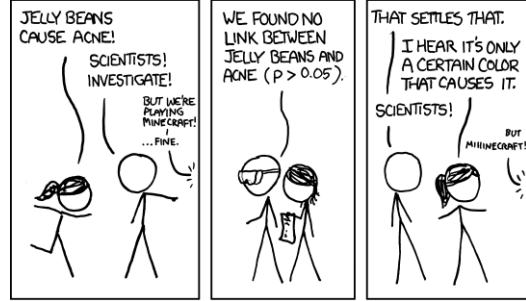
- Null: “All groups are the same”.
- Alternative: “**At least 1 group** is different”.

Performing Multiple Hypothesis Tests Anyways

- > For non-ANOVA methods, remember that performing many hypothesis tests increases our risk of incorrectly rejecting one of the null-hypotheses.
- > To compensate for this we decrease the p-value cutoff.
- > The most common way of doing this is with the Bonferroni Correction.

$$p' = \frac{p}{(\# \text{ of Hypotheses})}$$

- So if we chose $p=0.05$ as a cutoff. If we test 20 Hypotheses, each one should be tested at a significance level of $0.05/20 = 0.0025$.
- > This correction is argued to be too strong and other approximations for a new-p can be used instead.
 - Tukey's Range Test
- > This is VERY important in genetics/bioinformatics.
 - If we have thousands of genes we are testing the significance of or between, then we have to correct for this.



A caution on P-values

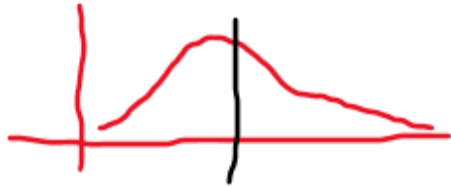
- > P-values have been over-glorified and grossly abused in science, statistics, the media, journals, etc...
- > "P-value Fishing"- the repeated testing of hypotheses until an interesting result is found.

ASA on P-values:

1. P-values can indicate how incompatible the data are with a model or hypothesis.
2. P-values don't measure the probability that a model or hypothesis is true or was observed by random chance.
3. Scientific conclusions and business/policy decisions shouldn't be based only on a p-value.
4. Proper Inference requires full reporting and transparency.
5. P-values don't measure the effect size or result importance.
6. By itself, a p-value doesn't provide a good measure of evidence regarding whether a model/hypothesis is true.

Using Simulations for Custom Hypothesis Test

- > Sometimes we calculate a statistic of interest that we don't know the distribution it should be.
- > To estimate this distribution, we can use simulation/sampling procedures (like bootstrapping) to find the distribution.
 - Then do the statistic calculation on your specific sample to get a p-value.
- > E.g. Calculating the difference between the median salaries between two different populations.



Median salary of “Software Developers”=med1



Median salary of “Software Engineers”=med2

- > $(\text{med1} - \text{med2})^2$ is distributed as?
 - Perform bootstrapping on the NULL hypothesis for both populations.
 - > 1. Sample from pop1 multiple times with replacement → Calculate distribution of median differences.
 - > 2. Where on that median-difference distribution pop2 median difference lies.
 - > (Also can do KS test here.)

Power of a Hypothesis Test

- > Statistically, every hypothesis test has a “power”.
- > The power is defined as $P(\text{rejecting Null} \mid \text{Null is False})$.
- > High power tests can easily find evidence enough to reject null hypothesis.
 - High power tests also tend to make more assumptions- it is easier to break assumptions when there are more of them.
- > Low power tests need more evidence to reject null hypothesis.
- > Example:
 - Doing pairwise tests of multiple groups (t-test) has less power than ANOVA.
 - > Because ANOVA only makes the assumption that “At least one group is different”
 - > Pairwise t-tests makes more assumptions: “Group A different from B”, “Group B different from C”...
- > The Power also depends on:
 - # samples in dataset.
 - Variability in dataset.
- > Calculating the power can be difficult analytically. Often we just simulate the test many times to get an idea of the Power.
 - Note: `scipy.stats` HAS some of the analytical solutions for common tests built in. (see in lab)

notebook time

we return to the lecture later

the end