3251 Foundations of Data Science

Data Science Fundamentals Certificate

Module 4

INFERENCE

Course Roadmap

Module / Week	Title
1	Introduction to Statistics for Data Science
2	Probability
3	Distribution of Random Variables
4	Inference Part 1
5	Inference Part 2
6	Linear Regression
7	Multiple Linear Regression
8	Logistic Regression
9	Introduction to Bayesian Inference
10	Multi-level Models
11	Markov Chain Monte Carlo
12	Presentations
13	Final Exam

Module 4: Learning Objectives

- Review of Inferential Statistics
- Hypothesis Testing
  - Point Estimate
  - Confidence Interval
  - Significance Level
  - P-Value
  - Z-Score
  - T-Distribution/T-Statistic
- Paired Data

Key Topic Overview – Inference

- Hypothesis Testing
- The t-Distribution
- Paired Data
- Difference Between Two Means
- Inference for Categorical Data

Required/Recommended Readings

Required:

1) OpenIntro, 3rd edition, by David Diez, Christopher Barr, Mine Centinkaya-Rundel, Copyright © 2015 OpenIntro.org

Recommended:

1) Think Stats: Probability & Statistics for Programmes Version 1.6.0, by Allen Downey, Copyright © 2011 Green Tea Press

2) Using R for introductory statistics, John Versani, Chapman and Hall/CRC Press

Section sub-header

REVIEW OF INFERENTIAL STATISTICS

Review of Inferential Statistics

- There are two types of statistics:
  - Descriptive statistics
  - Inferential statistics
- Descriptive statistics help a statistician to visualize what the data was showing, especially if there was a lot of it.
- Descriptive statistics do not make conclusions beyond the data we have or reach conclusions regarding any hypotheses we might have made.
- Typical measures include central tendency (mean, median, mode) and spread (variance, standard deviation, quantiles, range, etc.)

Review of Inferential Statistics (Con't)

- A statistician also tries to make statistical inferences about populations based on samples taken from the population.
- For example, a car dealer advertises that its new small truck gets 35 miles per gallon on the average.
- A tutoring service claims that its method of tutoring helps 90% of its students get an A or a B.
- A statistician will make a decision about these claims.
- The process involves the statistician collecting data from a sample and evaluating the data in order to make a decision as to whether or not the data supports the claim that is made about the population.

Review of Inferential Statistics (Con't)

- In summary, inferential statistics assume we do not have access to the entire population but a sample of it.
- And we would like to make generalization about the population from which the data is drawn.
- Because of sampling errors, a sample is not expected to perfectly represent the population
- Therefore, inferential statistics are concerned with
  - Testing of hypothesis to make a decision about a parameter
  - Estimation of parameters such as confidence levels

Section sub-header

HYPOTHESIS TESTING: SINGLE MEAN

Hypothesis Testing

- A statistician makes statistical inferences about a population based on samples taken from the population.
- Standard distributions are used to construct the model that produces data similar to the (real) data that has been collected
  - Same sample space
  - Similarly distributed
- The idea is to make an assumption on the model and then use the data collected to estimate the parameters of the model
- After the parameters are calculated, the statistician can use the model simulate a real world phenomenon or a business situation
- A statistician needs to
  - Understand the quality of the parameters for the model
  - Make valid statements about a population based on observations, i.e. test hypotheses and measure the weight of the evidence for/against a hypothesis

Hypothesis Testing (Con't)

- The sample data helps us to make an estimate of a population parameter.
  - Suppose you want to determine the average rent of a two-bedroom apartment. You might look in the classified section of the newspaper, write down several rents listed, and average them together. The resulting calculation gives a point estimate of the true mean.
- To make a statistical inference is to make a decision about a parameter
  - Suppose a car dealer advertises that its new small truck gets 35 miles per gallon, on the average. You might want to verify this claim by collecting data from a sample and evaluating the data. Then, you can make a decision as to whether or not the data supports the claim about the population. This process is called hypothesis testing.

Sample Estimates

- Particular distributions are associated with hypothesis testing
- For the normal distribution, the best estimate of the population parameters are the sample mean $\bar{x}$ and standard deviation s_x
- Every sample space we take would have a little different mean and standard deviation so the resulting model would be a little different, never quite the true μ and σ
- These measurements on a sample $\bar{x}$ and s_x are called point estimates

$$\bar{x} = \frac{\sum_{n} x_n}{n} \qquad s_x = \sqrt{\frac{\sum_{n} (x_n - \bar{x})^2}{n-1}}$$

Point Estimates

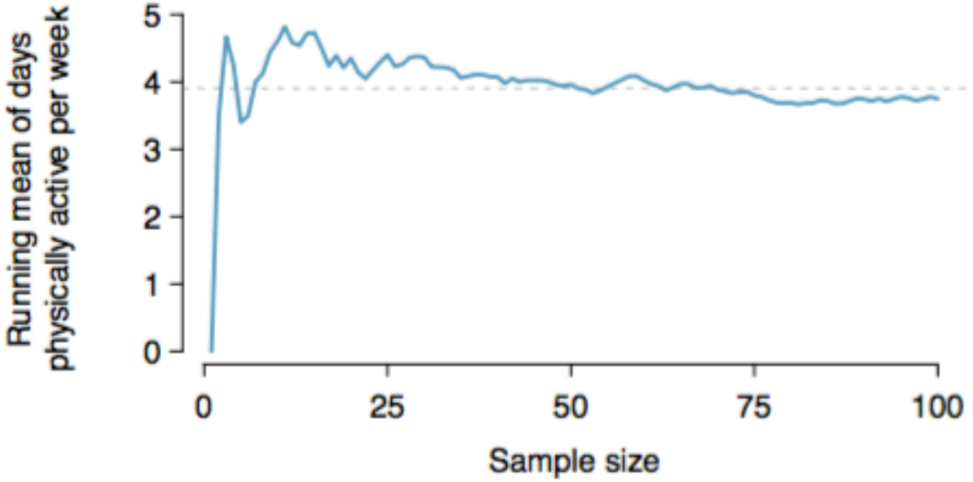


Figure 4.6: The mean computed after adding each individual to the sample. The mean tends to approach the true population average as more data become available.

- Sample estimates approximate the true population value

Hands On Demo

- Let's sample a normal distribution and take some point estimates on the mean and standard deviation
- Our point estimates (such as mean $\bar{x}$ and standard deviation s_x) show some random variation from sample to sample.
- Larger samples give point estimates with less variation: why do we expect this?

Sample Estimates (Con't)

- Assumptions when we perform hypothesis testing of a single population mean μ using a normal distribution
  - Take a simple random sample from the population
  - The population is normally distributed or the sample size is larger than 30 or both
  - The value of the population standard deviation σ is known

Sample Estimates (Con't)

- If we take a sample out of a population, how much confidence can we have in the point estimates as the parameters for our model
  - How can we quantify how good the point estimates are?
  - What can we say about the model if we want to use it to make decisions
- Since our point estimates move around just like any other random variable, we can measure the standard deviation of our point estimate sample means $\bar{x}$ which we denote as standard error $SE_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ given n independent observations
- We can consider the samples independent if it is a random sample of less than 10% of the population

Confidence Interval

- After calculating the point estimate, we construct a confidence interval in which the model parameter is believed to lie
- By applying the CLT, we can say that if a sample consists of at least 30 independent observations and the data are not strongly skewed, then the distribution of the sample mean is well approximated by the normal distribution
- If the population is skewed, has multiple peaks and outliers, we will need more samples
- If the sample size is fewer than 30, we will need to adjust for the uncertainty
- The empirical rule of normal distribution states that approximately 95% of the sample mean $\overline{x}$ will be within two standard deviations (+/- 2 standard deviation) of the population mean μ, i.e. we are 95% certain the true value of the parameter for the population μ as a whole is within this range

Confidence Interval (Con't)

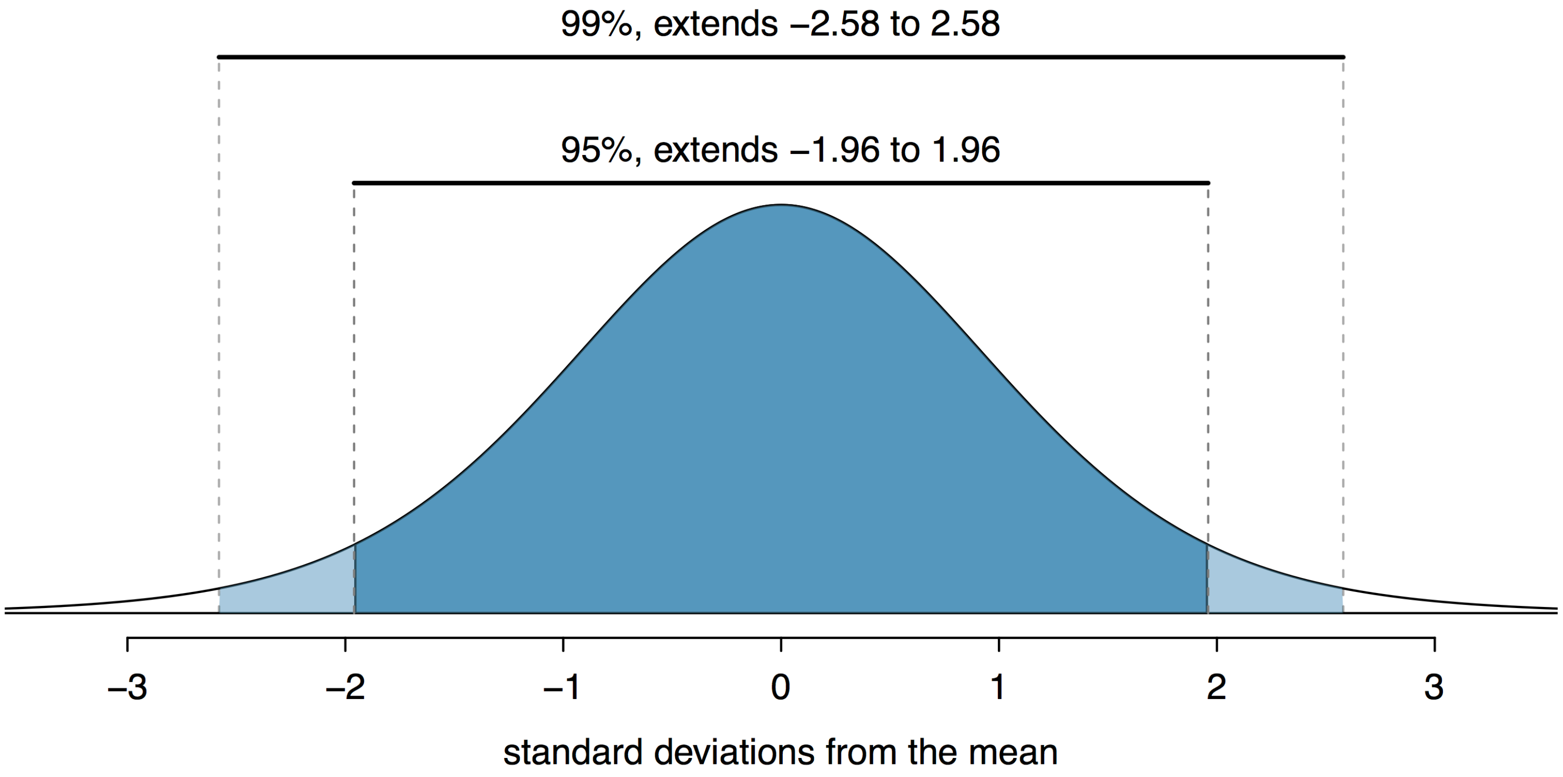


Figure 4.10: The area between -$z^\star$ and $z^\star$ increases as $|z^\star|$ becomes larger. If the confidence level is 99%, we choose $z^\star$ such that 99% of the normal curve is between -$z^\star$ and $z^\star$, which corresponds to 0.5% in the lower tail and 0.5% in the upper tail: $z^\star = 2.58$.

Confidence Interval (Con't)

- Suppose we do not know the population mean μ but we do know that the population standard deviation is $\sigma = 1$ and our sample size is 100.
- Then by the Central Limit Theorem, the standard deviation for the sample mean is $$\frac{\sigma}{\sqrt{n}} = \frac{1}{\sqrt{100}} = 0.1$$
- The sample mean $\bar{x}$ is within 0.2 units of μ.
- In other words, μ is between $\bar{x} - 0.2$ and $\bar{x} + 0.2$ in 95% of all the samples.
- Suppose the sample produced a sample mean $\bar{x} = 2$. Then the unknown population mean μ is between 1.8 and 2.2.
- The 95% CI for the unknown population mean μ is (1.8, 2.2).

Hypothesis Testing (Con't)

- There are 2 types of hypothesis testing:
  - Single mean
  - 2 means
- First, we will look at hypothesis testing of a single mean.
- Then we will look at hypothesis testing of paired data having two means.

Hypothesis Testing (Con't)

- The following are examples of hypothesis testing of a single mean:
  - We want to test whether the average GPA in US colleges is 2.0 (out of 4.0) or not.
  - We want to test whether 30% or less of the registered voters in Santa Clara County voted in the primary election.
  - We want to test if college students take less than five years to graduate from college, on the average.

Hypothesis Testing (Con’t)

- The goal of hypothesis testing is to answer the question, “Given a sample and an apparent effect, what is the probability of seeing such an effect by chance?”
- The actual test begins by considering two hypotheses the null hypothesis H_0 and the alternate hypothesis H_A.
- The null hypothesis states that there is no difference between the 2 claims being compared (or … no relationship between 2 phenomena being measured …, or … no association between 2 groups …).
- We assume the null hypothesis is true i.e. there is no difference unless there is evidence “strong enough” to reject it.

Hypothesis Testing (Con't)

- Therefore, hypothesis testing examines the sample data to determine which of H_0 or is true
  - The null hypothesis H_0: it is a statement about the population that will be assumed to be true unless it can be shown to be incorrect beyond a reasonable doubt.
  - The alternate hypothesis H_A: it is a claim about the population that is contradictory to H_0 and what we conclude when we reject H_0.
- We can either:
  - "reject H_0 in favor of H_A" if the sample information favors the alternate hypothesis or
  - "do not reject H_0" if the sample information favors the null hypothesis, meaning that there is not enough information to reject the null hypothesis

Hypothesis Testing (Con’t)

- There are four possible outcomes, as summarized in the following table:

		Test conclusion	
		Do not reject H_0	Reject H_0 in favour of H_A
Truth	H_0 true	√	Type 1 Error
	H_A true	Type 2 Error	√

- The evidence to decide which hypothesis to pick is in the form of sample data
- We will next look at how to estimate sample distributions

P-Value

- Let's calculate the actual probability of getting the test result, called the p-value
- The p-value is the probability that an outcome of the data (for example, the sample mean) will happen purely by chance when the null hypothesis H_0 is true
- A large p-value calculated from the data indicates that the sample result is likely happening purely by chance; therefore, the data support the null hypothesis so we do not reject it
- In other words, the p-value is the probability of observing data at least as favorable to H_A from our current data set, if the null hypothesis is true

P-Value (Con’t)

Suppose a baker claims that his bread height is more than 15 cm, on the average. Several of his customers do not believe him. To persuade his customers that he is right, the baker decides to do a hypothesis test. He bakes 10 loaves of bread. The average height of the sample loaves is 17 cm. The baker knows from baking hundreds of loaves of bread that the standard deviation for the height is 0.5 cm.

Ho: $\mu \leq 15$
Ha: $\mu > 15$.

The distribution for the test is normal with mean $\mu = 15$ and
standard deviation $\sqrt{\sigma/n} = \sqrt{0.5/10} = 0.16$.

The p-value, then, is the probability that a sample average is the same or greater than 17 cm when the population mean is, in fact, 15 cm.

p-value = P (sample average > 17) ~ 0

Because the outcome of 17 cm is so unlikely (meaning it is happening NOT by chance alone), we conclude that the evidence is strongly against the null

P-Value (Con’t)

- We compare the p-value and a preset value α (also called a significance level) to decide whether or not to reject the null hypothesis
- α is the probability that the effect (or relation or association) that we want to assess is due to chance alone or in other words, the probability of a Type I error (rejecting the null hypothesis when the null hypothesis is true).
- α may or may not be given to you at the beginning of the problem.
- When you make a decision to reject or not reject Ho, do as follows:
  - If α > p-value, reject H_0. The results of the sample data are significant to provide sufficient evidence to conclude that H_0 is an incorrect belief and that the alternative hypothesis H_A may be correct
  - If $\alpha \leq$ p-value, do not reject H_0. The results of the sample data are not significant to provide sufficient evidence to conclude that the alternative hypothesis H_A may be correct
  - When we "do not reject H_0", it does not mean that we should believe that H_0 is true. It simply means that the sample data has failed to provide sufficient evidence to cast serious doubt about the truthfulness of H_0

P-Value (Con’t)

- α is related to the confidence interval CI as

  $\alpha = 1 - CI$

- If for example we want to be 95% confident, then

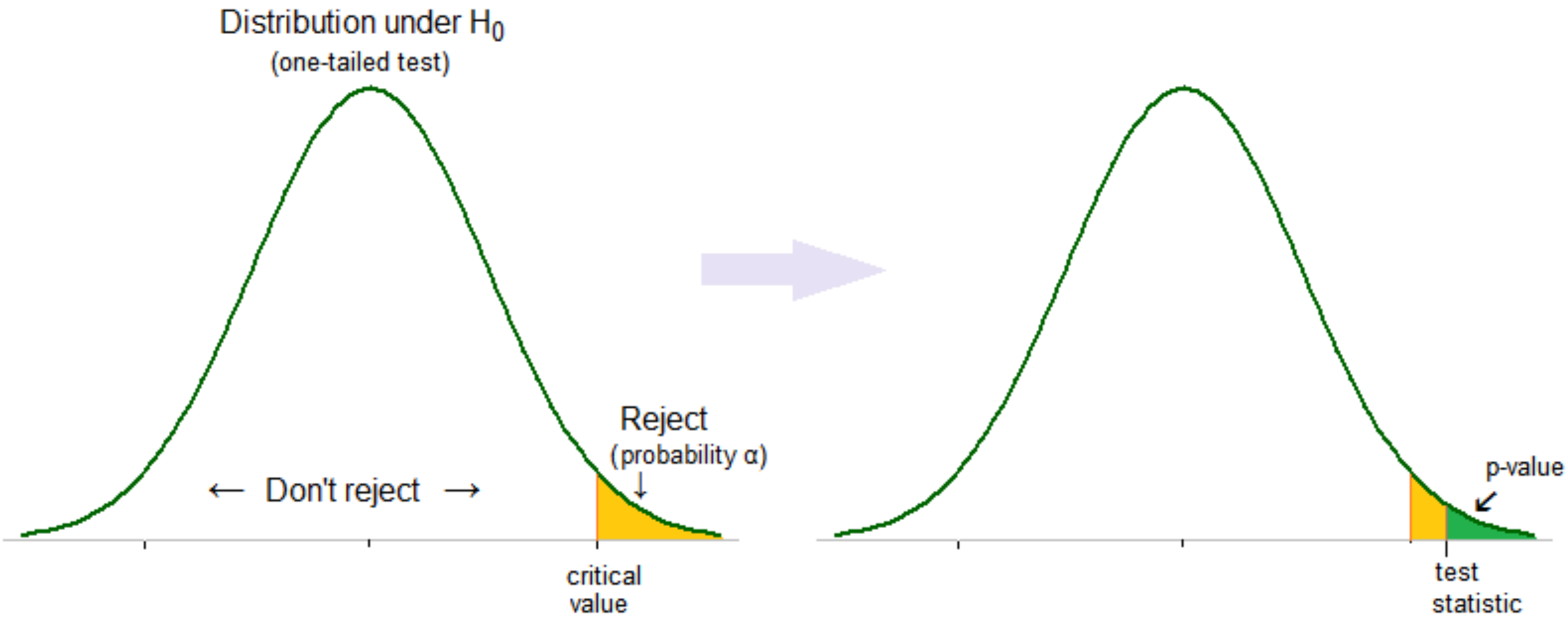
  $\alpha = 1 - 0.95 = 0.05$

- α may or may not be given at the beginning of the problem; if no level of significance is given, it is most often set at 0.05

P-Value (Con't)

- The p-value can in the left tail, the right tail, or split evenly between the two tails. For this reason, we call the hypothesis test left, right, or two tailed
- The alternate hypothesis H_A tells us if the test is left, right, or two-tailed. It is the key to conducting the appropriate test.

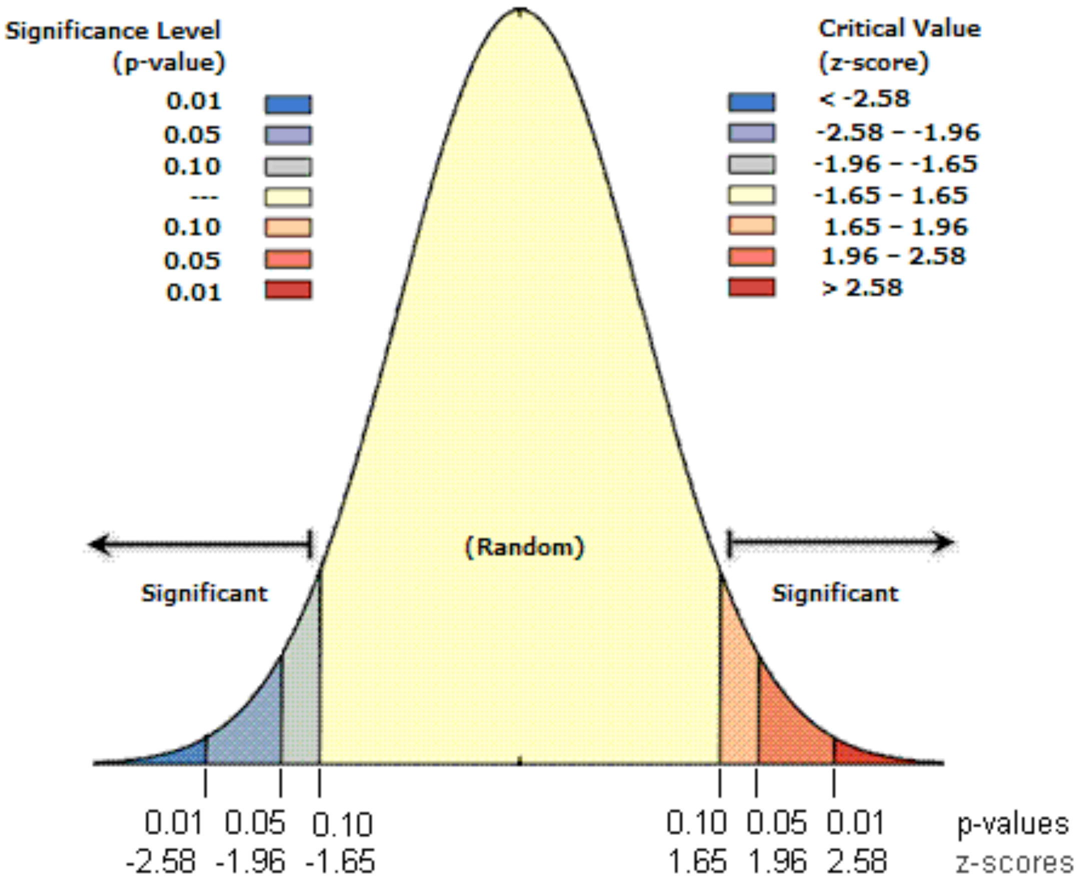
P-Value (Con't)


Distribution under H_0
(one-tailed test)
Reject
(probability α)
← Don't reject →
critical value
p-value
test statistic

Z-Score

- The z-score is the number of standard deviations an observation is away from the mean
- A.k.a. standard scores, z-values, normal scores
- A positive z-score means the observation is above the mean and a negative one below
- A z-score can be calculated from $z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$.
- From the z-score, the p-value can be calculated from the standard normal distribution table http://math.arizona.edu/~rsims/ma464/standardnormaltable.pdf or z-score calculator

P-Value and Z-score



http://pro.arcgis.com/en/pro-app/tool-reference/spatial-statistics/what-is-a-z-score-what-is-a-p-value.htm

The t-Distribution

- According to the central limit theorem (CLT), the sampling distribution (e.g. sample mean) of a large number of independent, identically distributed (IID) random variables will be approximately normal, regardless of the underlying distribution.
- Therefore, when we know the standard deviation of the population, we can compute a z-score (the z-score indicates how many standard deviations the sample mean is from the true mean) and use the normal distribution to evaluate probabilities with the sample mean and variance.
- There are two problems with this approach:
  - Sample sizes are sometimes small, and
  - Often we do not know the standard deviation of the population

The t-Distribution (Con't)

- The t-distribution (aka Student's t-distribution) is used to estimate population parameters when the sample size is small and/or when the population variance/standard deviation is unknown.
- When either of these problems occur, statisticians rely on the distribution of the t-statistic (also known as the t-score), whose values are given by:

$$t = \frac{\bar{x} - \mu}{\frac{s_x}{\sqrt{n}}} \qquad \bar{x} = \frac{\sum_n x_n}{n} \qquad s_x = \sqrt{\frac{\sum_n (x_n - \bar{x})^2}{n-1}}$$

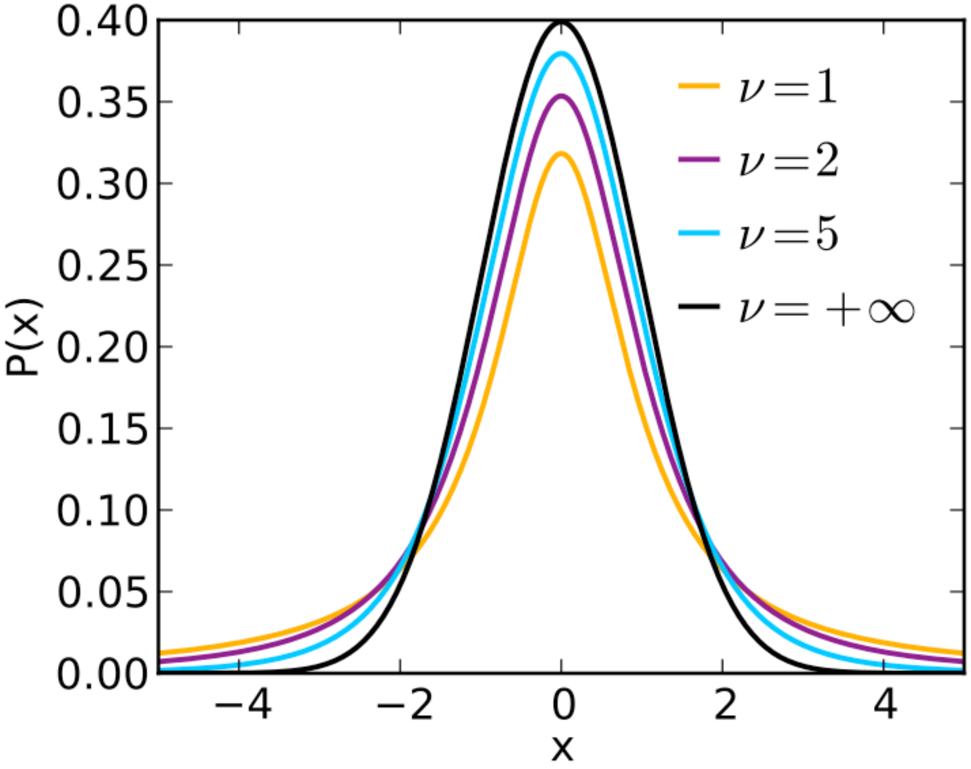
where $\bar{x}$ is the sample mean, μ is the population mean, s_x is the standard deviation of the sample, and n is the sample size.

- The distribution of the *t* statistic is called the t-distribution or the Student t-distribution.

The t-Distribution (Con't)

- The t-distribution can be used with any statistic having a bell-shaped distribution (i.e., approximately normal).
- The sampling distribution of a statistic should be bell-shaped if any of the following conditions apply.
  - The population distribution is normal.
  - The population distribution is symmetric, unimodal, without outliers, and the sample size is at least 30.
  - The population distribution is moderately skewed, unimodal, without outliers, and the sample size is at least 40.
  - The sample size is greater than 40, without outliers.
- The t-distribution should NOT be used with small samples from populations that are NOT approximately normal.

The t-Distribution (Con’t)



https://en.wikipedia.org/wiki/Student's_t-distribution

Statistician William Sealy Gosset, known as "Student"

The t-Distribution (Con’t)

- When a sample of size n is drawn from a population having a normal (or nearly normal) distribution, the sample mean can be transformed into a t-statistic $t = \dfrac{\overline{x} - \mu}{\frac{s_x}{\sqrt{n}}}$ with $n-1$ degrees of freedom.
- t_α presents the t-statistic that has a cumulative probability of $(1-\alpha)$ with significant level α .
- Warning: if the population is NOT normal and the sample size is small, then we CANNOT use either the z-score and the t-statistic

The t-Distribution (Con’t)

- Statistics students believe that the average score on the first statistics test is 65. A statistics instructor thinks the average score is higher than 65. He samples ten statistics students and obtains the scores 65; 65; 70; 67; 66; 63; 63; 68; 72; 71. He performs a hypothesis test using a 5% level of significance.
- Ho: μ = 65 Ha: μ > 65
- Since the instructor thinks the average score is higher. This means the test is right-tailed.
- You are only given n = 10 sample data values. Notice also that the data come from a normal distribution. This means that the distribution for the test is a t-distribution.
- Use t_{df}. Therefore, the distribution for the test is t_9 where n = 10 and df = 10 − 1 = 9.
- Sample mean and sample standard deviation are calculated as 67 and 3.1972 from the data respectively.
- t-statistic = 1.978141 using the formula above in the slide
- p-value = P(X > 67) = 0.0396 calculated using the p-value from t-statistic calculator
- Since α = 0.05 and p-value = 0.0396. Therefore, α > p-value.
- Since α > p-value, reject H_0.
- This means you reject μ = 65. In other words, you believe the average test score is more than 65.
- At a 5% level of significance, the sample data show sufficient evidence that the mean (average) test score is more than 65, just as the instructor thinks.

Margin of Error

- Depending on whether you know the population standard deviation or not,
  - Margin of error is calculated as critical z value x standard deviation
  - Margin of error is calculated as critical z value x standard error
- When the sampling distribution is nearly normal, the critical value can either be expressed as the z-score or *t*-statistic
  - Compute the significant level α
  - Find the critical probability $p^* = 1 - \frac{\alpha}{2}$
  - Express the critical value as the
    - z-score having a cumulative probability equal to p^* or
    - t-statistic after calculating the degrees of freedom (DF=n-1 with n samples) and having a cumulative probability equal to p^*.

Summary

- We now have a general framework for statistically testing hypotheses once we've selected a model and estimated its parameters
- Hypothesis testing boils down to determining how unusual our result would be if our skeptical nothing-to-see-here hypothesis were true
- We can quantify just how unusual it is using p-value
- We can quantify how good our model is using standard error

Summary (Con’t)

- Hypothesis testing for nearly normal point estimates
  - First write the hypotheses in plain language, then set them up in mathematical notation
  - Identify an appropriate point estimate of the parameter of interest
  - Verify conditions to ensure the standard error estimate is reasonable and the point estimate is nearly normal and unbiased
  - Compute the standard error. Draw a picture depicting the distribution of the estimate under the idea that H_0 is true. Shade areas representing the p-value.
  - Using the picture and normal model, compute the test statistic (z score) and identify the p-value to evaluate the hypotheses. Write a conclusion in plain language.
  - http://xkcd.com/882

Source: OpenIntro Statistics, 3rd Ed., p. 199

Section sub-header

HYPOTHESIS TESTING: PAIRED DATA, TWO MEANS

Hypothesis Testing: Paired Data, Two Means

- Previously we have learnt how to conduct hypothesis tests on single means
- Next we want to extend what we have learnt about hypothesis testing to comparing numerical datasets
- We turn our attention to comparing the results of two or more experiments (possibly with different numbers of observations of each)
- For example, researchers are interested in the effect aspirin has in preventing heart attacks.
- Typically, one group is given aspirin and the other group is given a placebo. Then, the heart attack rate is studied over several years by comparing the 2 groups.
- The general procedure is still the same, just expanded.

Hypothesis Testing: Paired Data, Difference Btwn 2 Means (Con’t)

- When comparing experimental outcomes, we want to answer the question whether the datasets for the two experiments are independent?
- Let’s review independence: http://en.wikipedia.org/wiki/Correlation_and_dependence
- The procedures are different depending on whether the data are independent or paired

Hypothesis Testing: Paired Data, Two Means (Con't)

- To compare 2 averages, we work with 2 groups.
- The groups are classified either as independent or matched pairs.
  - Independent groups mean that the 2 samples taken are independent, that is, sample values selected from one population are not related in any way to sample values selected from the other population.
  - Matched pairs consist of 2 samples that are dependent.
- We will learn how to deal with both paired and independent data

Hypothesis Testing: Paired Data, Two Means (Con't)

- First we look at comparing 2 independent population means with unknown population standard deviations
  - The two independent samples are simple random samples from two distinct populations
  - Both populations are normally distributed with the population means and standard deviations unknown unless the sample sizes are greater than 30.
  - If the sample sizes are greater than 30, the populations need not be normally distributed

Hypothesis Testing: Paired Data, Two Means (Con’t)

- A difference between the two samples depends on both the means and the standard deviations.
- Very different means can occur by chance if there is great variation among the individual samples.
- In order to account for the variation, we take the difference of the sample means $\bar{x}_1 - \bar{x}_2$ and divide by the standard error in order to standardize the difference. The result is a t-score test statistic
- The standard error is $$\sqrt{\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}}$$

where s_1 and s_2 are sample standard deviations

Hypothesis Testing: Paired Data, Two Means (Con't)

- The t-statistic is $\dfrac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{\sqrt{\dfrac{(s_1)^2}{n_1}+\dfrac{(s_2)^2}{n_2}}}$ and approximated by

  the t-distribution with degrees of freedom

$$df=\frac{\left(\dfrac{(s_1)^2}{n_1}+\dfrac{(s_2)^2}{n_2}\right)^2}{\dfrac{1}{n_1-1}\left(\dfrac{(s_1)^2}{n_1}\right)^2+\dfrac{1}{n_2-1}\left(\dfrac{(s_2)^2}{n_2}\right)^2}$$

- When both n_1 and n_2 are larger than 5, the approximation is very good

Hypothesis Testing: Paired Data, Two Means (Con't)

- Next we look at comparing 2 independent population means with known population standard deviations
  - This scenario is unlikely
  - Replace s_1 and s_2 with σ_1 and σ_2
  - Same equations as the case with unknown population standard deviations

Section sub-header

HYPOTHESIS TESTING: DIFFERENCE BETWEEN TWO MEANS

Hypothesis Testing: Paired Data, Difference Between 2 Means

- Finally we look at comparing matched or paired samples
- In a hypothesis test for matched or paired samples, subjects are matched in pairs and differences are calculated
- The differences are the data
- The population mean for the differences μ_d is then tested using a Student-t test for a single population mean with n − 1 degrees of freedom where n is the number of differences.

Hypothesis Testing: Paired Data, Difference Btwn 2 Means (Con't)

- The t-statistic is $\dfrac{\bar{x}_d - \mu_d}{\dfrac{s_d}{\sqrt{n}}}$
- Assumptions:
  - 2 samples are drawn from the same set of objects
  - Simple random sampling is used
  - The samples are dependent
  - Differences are calculated from the matched or paired samples
  - The differences form the sample that is used for the hypothesis test
  - The matched pairs have differences that either come from a population that is normal or the number of differences is greater than 30 or both

Hypothesis testing

- A study was conducted to investigate the effectiveness of hypnotism in reducing pain.

Subject	A	B	C	D	E	F	G	H
Before	6.6	6.5	9.0	10.3	11.3	8.1	6.3	11.6
After	6.8	2.4	7.4	8.5	8.1	6.1	3.4	2.0
Diff	0.2	-4.1	-1.6	-1.8	-2.3	-2.0	-2.9	-9.6

- Are the sensory measurements, on average, lower after hypnotism? Test at a 5% significance level.
- The sample mean and sample standard deviation of the differences are:
  $\text{diff}_{\text{mean}} = -3.13$ and $s_d = 2.91$
- $H_0 : \mu_d \geq 0$; there is no improvement. μ_d is the population mean of the differences.
- $H_a : \mu_d < 0$; there is improvement
- $df = n - 1 = 8 - 1 = 7$
- Find t_7, calculate the p-value using the Student-t distribution: p-value = 0.0095, $\alpha = 0.05$
  and
  Since $\alpha >$ p-value, reject H_0

Questions?