The University of Texas at Arlington

Basics of Statistics and Tests

CSE 5301 – Data Modeling and Analysis Techniques

Dr. Gergely Záruba

Probability vs. Statistics

- Probability and statistics are often blurred together by students. They are related but accomplish two completely different tasks.
  - Statistics
    - A field of applied mathematics. Used in other applied fields to draw conclusions about data gathered.
    - Used for collection, analysis, and interpretation of data.
    - Thus, the analysis part of our studies in this class.
  - Probability
    - A field of theoretical mathematics used in applied fields as a modeling tool.
    - Predicts the likelihood of events given a set of assumptions (e.g. a distribution)
    - Thus, the modeling part of our studies in this class.

2

Statistics

- Statistics deals with real world data
  - Data collection
    - How do we have to collect the data to get valid results?
  - Analysis of data
    - What properties does the data have?
    - How do we ensure our data analysis is valid?
    - What distribution does the data come from?
  - Interpretation of the data
    - Is the data completely different from other data?
    - What could cause issues in the data?

3

STATISTICS

4

Statistics

- Statistics deal wit interpreting and analyzing data collected.
- Statistics attempt to represent the important characteristics of a set of data items (or of a probability distribution) and the uncertainty contained in the set (or the distribution).
  - Statistics represent different attributes of the probability distribution represented by the data
  - Statistics are aimed at making it possible to analyze the data based on its important characteristics

5

Experiment and Sample Space

- A (random) experiment is a procedure that has a number of possible outcomes and it is not certain which one will occur
- The sample space is the set of all possible outcomes of an experiment (often denoted by S).
  - Examples:
    - Coin : S={H, T}
    - Two coins: S={HH, HT, TH, TT}
    - Lifetime of a system: S={0..∞}

6

Mean

- The sample mean represents the average value of data set $X=\{x_1,x_2,\ldots x_N\}$

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

- The sample mean is the expected value of a statistical random variable, i.e., the expected value of a data item drawn at random

$$m = E[X]$$

7

Median and Mode

- The median is the middle element of the sample set when ordering them by their magnitude
  - If the number of samples is odd this is definite if the number of samples is even then it is usually defined as the average of the two middle values.
- The mode of a sample set is the most frequently obtained (i.e. most likely) value
  - Sample sets may not contain a mode (no samples are drawn twice), or may contain several modes.

8

Variance and Standard Deviation

- The variance provides a scalar representing the spread of the data set. In a data set $X=\{x_1,x_2,\ldots x_N\}$ an **unbiased estimate** s_u^2 for the variance can be calculated as

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

  - $N-1$ is often called the number of degrees of freedom of the data set
- The standard deviation s is the square root of the variance

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

  - In the case of a sample set, s is often referred to as standard error
  - Unlike the variance, the standard deviation is measured in the same units the original data was measured in.

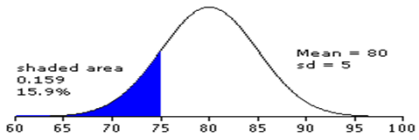
9

CONFIDENCE INTERVALS

10

Percentiles

- To determine the likelihood that a data item could come from a distribution we have to be able to determine percentiles
  - A data item belongs to the n^{th} percentile if the likelihood to obtain a value that is equal to the data item or even further away from the distribution mean is greater or equal to $n\%$



  - For certain distributions percentiles can be relatively easily calculated for other we usually use tables.

11

Percentiles in Normal Distributions

- The percentile in a normal distribution is a function of the distance of the data value from the mean and of the standard deviation

$$z = \frac{X - \mu}{\sigma}$$

z	Area from −∞ to z
-3.0	.0013
-2.5	.0062
-2.0	.0227
-1.5	.0668
-1.0	.1587
-0.5	.3085
0.0	.5000
0.5	.6915
1.0	.8413
1.5	.9332
2.0	.9772
2.5	.9938
3.0	.9987

  - E.g. a data value that is more than 1.5 standard deviations larger than the mean of the distribution occurs only with probability 0.0668

12

Percentiles of Distributions

- If the distribution of the population is normal, the z-value and the z-table allow to compute how likely it would be to randomly draw the particular data value (or one even further from the mean)
  - If the likelihood is not very small, then we should not assume that the data value is significant different from the value of the distribution
- Percentiles for general, skewed distributions are difficult to derive
  - So we will need to attempt to formulate our tests on statistics for which the distribution is approximately normal

13

Sampling Distributions

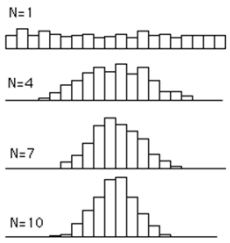
- Sample Distribution: The probability distribution underlying the random variable X from which individual data items come from (μ_X, σ_X)
- Sampling Distribution: The probability distribution of a statistic calculated from a set of randomly drawn actual data items (samples) (μ_S, σ_S)
  - Sampling Distribution of the mean: The distribution of the means of random data samples of size N. For a sample distribution with mean μ_X and standard deviation σ_X the mean μ_S and standard deviation σ_S of the sampling distribution of the mean over N samples is:

$$\mu_S = \mu_X \text{ and } \sigma_S = \frac{\sigma_X}{\sqrt{N}}$$

14

Central Limit Theorem

- For any sample distribution with mean μ_X and standard deviation σ_X, the sampling distribution of the mean approaches a normal distribution with mean μ_X and standard deviation $\frac{\sigma_X}{\sqrt{N}}$ as N becomes large.
  - This means that percentiles for the sampling distribution of the mean are easier to compute than for the sample distribution.



15

Confidence Interval for the Mean

- If we have a large (N≥30) set of samples $\{X_i\}$ coming from a distribution with finite (but unknown) mean μ_X and finite variance s_X, can we say how close the sample average $\bar{x}$ is to μ_x?
- We can have probabilistic guarantees on this (probabilistic bounds):

$$P(\bar{x} - e \leq \mu_X \leq \bar{x} + e) = 1 - \alpha$$

  - Where a is our significance level, 1-a is called the confidence level, and the interval $[\bar{x} - e, \bar{x} - e]$ is called the confidence interval.
- For the same set of samples, if the interval (i.e., e) is increased, our confidence level goes up, and vice-versa.

16

Confidence Interval for the Mean (2)

- $P(c_1 \leq \mu_X \leq c_2) = 1 - \alpha$, $c_1 = \bar{x} - e$, $c_2 = \bar{x} + e$
- What is the exact relationship between *e* and α?
- We know that the distribution of the sample average for large N-s is $N\left(\mu_X, \frac{\sigma_X^2}{N}\right)$
  - $\frac{\sigma_X}{\sqrt{N}}$ is called the standard error.
  - If N is large, we can estimate σ_X by s_X $(s_X \approx \sigma_X)$
- I.e., what is the probability that $\bar{x}$ falls within a dense part of $N\left(\mu_X, \frac{s_X^2}{N}\right)$ (thus integrating $N\left(\mu_X, \frac{s_X^2}{N}\right)$ between c_1 and c_2 ?
- Using a z-table thus: $\mathrm{e} = z\left(1 - \frac{\alpha}{2}\right) * \frac{s_X}{\sqrt{N}}$ where z(p) is the p^{th} percentile of the standard normal.

17

Confidence Interval for Fixed Samples

- If we have a fixed sample set (i.e., we cannot obtain more samples), then we have a choice:
  - Have a target confidence level and calculate the confidence interval for that. Target confidence levels in engineering start from 95%.
  - Have a target confidence interval (or e error) and calculate the significance (confidence) level for that.
    - Sometimes we define the relative error as $e_r = e/\bar{x}$. Target relative error levels of 5% or less are acceptable.
- If we have both a target confidence level and confidence interval, then the only thing we can do is to test if our samples abide that.
- When plotting our data, we can show our confidence intervals on each data (lazy people show the sample standard deviation), and we need to disclose N.

18

If More Samples Can Be Obtained

- If both *e* and α have target values (e_T and α_T) and we have the liberty to obtain an arbitrary number of samples, how many samples N_t should we get?
  1. We need to estimate N_T based on an initial sample set. Thus we need to obtain an initial *j=0* set of samples $\{X_i^{(0)}\}$ with N_0=30.
  2. Calculate of $\bar{x}^{(j)}$ and $s_X^{(j)}$, Evaluate $e^{(j)}$ for $\alpha^{(j)}$ on $\{X_i^{(j)}\}$. If test succeeds (if $\alpha^{(j)} < \alpha_T$ when $e^{(j)}=e_T$) then we are done.
  3. If test fails, then estimate N_{j+1}. $N_{j+1} = \left(z\left(1-\frac{\alpha}{2}\right) * \frac{s_X^{(j)}}{e_T} \right)^2$
  4. Throw away(!) $\{X_i^{(j)}\}$ and obtain a new set of N_{j+1} samples $\{X_i^{(j+1)}\}$; and *j:=j+1*. Go back to the second step.

19

Confidence Intervals – Remember!

- Confidence intervals based on "Z" are completely scientific to use when:
  - The distribution from which the samples come from is normal, the population variance is known, for any number of samples.
  - The distribution from which the samples come from is well behaved, the population variance is known, the number of samples is sufficiently large.
- Confidence intervals based on "Z" are completely accepted to use when:
  - The distribution from which the samples come from is well behaved, the population variance is unknown but is estimated by an unbiased estimator, the number of samples is sufficiently large.
- Confidence intervals based on "Z" are unaccepted:
  - In all other cases!

20

Student's t

- Confidence intervals can be also calculated in the case when:
  - The distribution from which the samples come is normal, the population variance is unknown, and the number of samples is less than sufficiently large.
- The question becomes, how can we introduce the error in our calculations that comes from not knowing the population variance but only the sample variance?
  - With the Z-based confidence intervals, we were able to use the CLT and say, that: $\frac{\bar{x}-\mu_X}{s_x/\sqrt{N}} = Z \approx N(0,1)$ for the distribution.
  - Is there a distribution that can model $\frac{\bar{x}-\mu_X}{s_x/\sqrt{N}}$ when N is not sufficiently large? There is a distribution, called the Student's t distribution. Student's t is "flatter than a normal". It actually is a set of distributions (or a distribution with an additional parameter) differing in the ***degrees of freedom*** (N-1).

Confidence Intervals – using "t"

- Use the Student's t distribution instead of z for confidence intervals if:
  - The distribution from which the samples come is normal, the population variance is unknown, and the number of samples is less than sufficiently large.
- Pay attention to look at the right degrees of freedom (always one less than the sample size N).
  - Using a t-table thus: $\mathrm{e} = t\left(1-\frac{\alpha}{2}, N-1\right) * \frac{s_x}{\sqrt{N}}$ where t(p,m) is the p^{th} percentile of the m-th degree of freedom Student's t.

HYPOTHESIS TESTING

Hypothesis Testing

- Hypothesis testing is a statistical method used to evaluate if a particular hypothesis about data resulting from an experiment is reasonable.
  - Uses statistics to represent the data
    - Value of the data
    - Distribution of the data
  - Determine how likely it is that a given hypothesis about the data is correct

Hypothesis Testing (2)

- Hypothesis testing can be used to determine if a particular hypothesis about a set of observations (data) could be trusted
- In general hypothesis testing can be:
  - One-sample set
    - We are trying to establish the relationship ("sameness" or a difference) between a sample set and a random variable.
  - Two-sample sets
    - We are trying to establish the relationship ("sameness" or a difference) between two sample sets.

25

One-sample set Hypothesis Testing

- We are trying to establish the relationship ("sameness" or a difference) between a sample set and a random variable.
- Example:
  - The average and standard deviation of the processing delay in a router manufactured by Riporouter are:

    μ_R=170ms σ_R=1.2ms
  - The competitors, Suprouter product was measured at ten data packets (with an average m_S=171.3):

    $x_i \in$ {167,174,168,180,173,182,160,170,174,165}
  - Can we conclude that the delay of this second router is on the average higher than the delay of Riporouter's product?

26

Two-sample set Hypothesis Testing

- We are trying to establish the relationship ("sameness" or a difference) between two sample sets.
- Example:
  - We measure the processing delay in a router manufactured by Riporouter as: $y_i \in$ {168,169, 173,170,173,182,160,170,174,165} (with an average m_R=170ms)
  - The competitors, Suprouter product was measured at ten data packets: $x_i \in$ {167,174,168,180,173,182,160,170,174,165} (with an average m_S=171.3)
  - Can we conclude that the delay of this second router is on the average higher than the delay of Riporouter's product?
  - Does it matter if the measurement values are paired?

27

Hypothesis Testing (3)

- To be able to trust in a hypothesis on statistical data we have to make sure that the data set could not be the result of random chance
  - In the first example the hypothesis would be:

    H: $\mu_s > \mu_R$
  - To determine if the hypothesis has a base we have to make sure that we do not accept it if the data could be the result of random chance
    - What is the likelihood that the data set could be obtained by randomly sampling 10 items from the Riporouter?

28

The Logic of Hypothesis Testing

- The goal of hypothesis testing is to establish the viability of a hypothesis about a parameter of the population (very often the mean)
- Recall our example: Is the Suprouter delay on the average higher than Riporouter's?
  - Riporouter characteristics are: μ_R=170ms σ_R=1.2ms
  - Suprouter was measured: {167,174,168,180,173,182,160,170,174,165}
- Define hypothesis (also called ***alternative hypothesis***): H_A: $\mu_S > \mu_R$
- Set up ***null hypothesis*** (i.e. the "opposite" of H) H_0: $\mu_S = \mu_R$
- Compute the percentile and thus the likelihood H_0
  - If H_0 has more than a small likelihood then the data does not significantly support H_A (since the data could also represent H_0)
  - What are small likelihoods? Usually thresholds (***significance levels***) of 5% or smaller are used.

29

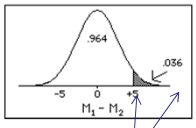
Rejecting the Null Hypothesis

- If the null hypothesis' likelihood is smaller than the significance level then the null hypothesis can be ***rejected***.
- Rejection implies that the null hypothesis is discarded in favor of the alternative hypothesis and the result is considered significant
  - Note, that a p-value less than 5% for the Null hypothesis does NOT imply a likelihood of 95% for the alternative hypothesis.
  - Note that it is NOT possible to show that the Null hypothesis is correct. Failure to reject the Null hypothesis does NOT imply acceptance of the Null hypothesis but rather that no significant conclusion could be drawn from the test
- There are four different situations, two resulting in correct decisions (when we accept or reject the null hypothesis correctly). However:
  - **Type-I error:** (α) When we reject the null hypothesis when it is actually true.
  - **Type-II error:** (β) When we do not reject (i.e., we accept) the null hypothesis when it actually is false.
  - **The power of a test:** is the probability that we reject H_0 when we H_A is true, which is 1-P(Type-II).

30

The Number of Tails

- Depending on the hypotheses we might be interested to know how the likelihood to generate data that is more extreme than the test data in a particular direction (e.g. the likelihood of it being larger than or equal to the given data) or in any direction (i.e. that it is further from the mean than the given data)
  - If we are only interested in data on one end of the distribution we perform a one-tailed test, i.e. we only count the percentile at one end of the distribution
  - If we are interested in both sides, we perform a two-tailed test which computes the percentile at both ends
  - If we are not sure we should choose a two-tailed test (which is more stringent)



Critical region

Critical regions

31

Which Tail Then?

- If you are looking for a decrease, i.e., H_A: …<... Then use one-tail, lower tail
- If you are looking for an increase, i.e., H_A: …>... Then use one-tail, upper tail
- If you are looking for a change i.e., H_A: …≠... Then use two tails.

- Always decide on the significance level first!

32

One-Sample set Z-test

- The Z Test is the most basic hypothesis test to evaluate a hypothesis related to a set of randomly drawn samples from an unknown distribution X ($\{X_i\}$ of size N with an average of: $\bar{x}$) and a probability distribution Y for which both the mean μ_Y and the standard deviation σ_Y are known. N needs to be sufficiently large. The question is: is X related to Y?
  - Z assumes that the sampling distribution of the means is normal
    - Either the sample distribution is normal or the sample size is very large (N≥30)
  - The alternate hypothesis could be:

    H_A: $\mu_X > \mu_Y$
  - Thus the null hypothesis is:

    H_0: $\mu_X = \mu_Y$
  - Compute z-value (the distance between $\bar{x}$ and μ_Y in standard error units)

    $$z = \frac{\bar{x} - \mu_y}{\sigma_Y / \sqrt{N}}$$ (the value on the domain of the standard normal)
  - Translate z-value to p-value and evaluate significance
    - We use a Z-table to determine the weight of the tail from z, i.e., (one half or the full) significance level p. If p is smaller than our significance goal a, then we reject the null hypothesis.. (e.g. p = 2.5% -> z=1.96)

33

What if I Know I am sampling a Normal?

- The previous procedure can be used for any number of samples (N does not need to be large) if we know the standard deviation of the random variable we are comparing the samples to.

34

What If σ_Y is Unknown?

- If the sample size is sufficiently large, it is generally accepted to estimate σ_Y with the standard deviation of the sample set s_X.

  $\sigma_Y \approx s_X$.
- Thus estimating the standard error by:

$$\frac{s_X}{\sqrt{N}}$$

- This is accepted as the second order statistics are expected to converge faster than the mean and as s_X is an overestimate of μ_X.

35

What If σ_Y is Unknown and N is Not Sufficiently Large?

- If we know that we are sampling from a normal, then instead of using z, we need to be using the proper degree of freedom Student's t when determining significance.
- Otherwise, these statistical methods cannot be used.

36

Two-sample Set Tests

- In some cases we need to test if two sets of samples $\{X_i\}$ and $\{Y_i\}$ come from the same distribution (or two distributions with the same mean). Here we need to know the two distributions' standard deviations. The sample sizes need to be sufficiently large.
- The null hypothesis is then:

  H_0: $\mu_X = \mu_Y$ which is: $\mu_X - \mu_Y = 0$

  Recall: $E[\bar{X} - \bar{Y}] = E[\bar{X}] - E[\bar{Y}]$ *and* $Var[\bar{X} - \bar{Y}] = Var[\bar{X}] + Var[\bar{Y}] = \frac{\sigma_X^2}{N_X} + \frac{\sigma_Y^2}{N_Y}$
- The alternate hypothesis can be one-sided or two-sided.
- z can be determined then as:

$$z = \frac{(\bar{x} - \bar{y}) - (\mu_x - \mu_y)}{\sqrt{\frac{\sigma_X^2}{N_X} + \frac{\sigma_Y^2}{N_Y}}}$$

- It gets simpler if the two variances are the same and even simpler if the two sample sizes are the same.

37

Unknown Variance, Two-Sample Sets

- a.k.a. two-sample pooled test
- If we can assume that σ_X and σ_Y are the same (although unknown), then:

$$z = \frac{(\bar{x} - \bar{y}) - (\mu_x - \mu_y)}{s_{XY}\sqrt{\frac{1}{N_X} + \frac{1}{N_Y}}}$$

where $s_{XY}^2 = \frac{s_X^2(N_X - 1) + s_Y^2(N_Y - 1)}{N_X + N_Y - 2}$

- If sample sizes are not large enough (but we assume that we sample from normal distributions) we need to use t statistics instead of z with the degrees of freedom of: $N_X + N_Y - 2$

38

Unknown Variances, Two-Sample Sets

- If we do not assume that σ_X and σ_Y are the same (although unknown), then:

$$z = \frac{(\bar{x} - \bar{y}) - (\mu_x - \mu_y)}{\sqrt{\frac{s_X^2}{N_X} + \frac{s_Y^2}{N_Y}}}$$

- If sample sizes are not large enough (but we assume that we are sampling from normal distributions) we need to use t statistics instead of z with the degrees of freedom of:

$$\upsilon = \left\lfloor \frac{\left(\frac{s_X^2}{N_X} + \frac{s_Y^2}{N_Y}\right)^2}{[(s_X^2/N_X)^2/(N_X - 1)] + [(s_Y^2/N_Y)^2/(N_Y - 1)]} \right\rfloor$$

39

Paired Tests

- Paired observations: we have two equal size sample sets, where samples were obtained under the same circumstances. Thus the pairing of the samples matters as they are not independent.
  - E.g.: Does algorithm 1 perform better than algorithm 2 based on their performance on a specific set of problems (the same problems for both)
  - A paired sample test avoids the variance caused by the controlled variable (e.g. the specific problem the algorithm is applied to) by establishing the sampling distribution over the differences in the value between paired data items from both sets:

$$\{D_i = X_i - Y_i\}$$

  - Example Hypotheses:

$$H_A: \mu_D > 0 \quad \text{thus:} \quad H_0: \mu_D = 0$$

  - Compute t-value (or z-value):

$$t_{n-1} = \frac{\bar{D} - \mu_D}{s_D/\sqrt{N}} = \frac{\bar{D}}{s_D/\sqrt{N}} \quad \text{(under the null hypothesis)}$$

  - Translate t-value to the corresponding p-value (as usual)

40

Two-Sample vs. Paired Samples

- The paired sample test is preferable whenever an additional variable is known which produces variations in the data items
  - Intuitively: a paired sample test often has smaller standard deviations because of the avoided variance in the conditional variables, thus should offer better confidences.
- If no conditional variable that would pair individual samples together is known to be relevant, the two-sample test is most of the time better because it uses more samples

41

TESTING OTHER STATISTICS

42

Statistics of "Other" Properties

- We have looked at setting a statistical relationship between the sample average and the mean of the underlying random variable.
- We could do this as we knew the distribution of the average (based on CLT for z or based on normals adding to normal for t).
- If we knew the sampling distribution of other statistics we could do the same for them as well…
- Variance (arguably the second most important value of a sample set) comes to mind…

43

Testing Variance

- Variance indicates the spread of the values drawn
  - Can correspond to "reliability", "noisiness", etc.
- To allow significance tests about hypotheses related to the variance we need to know what the sampling distribution of the variance of a sample distribution is and be able to compute its percentiles
  - If we know this distribution, significance tests can be performed in the same way as for the mean
- The variance of a random variable can be interpreted itself as a random variable $(x-\mu)^2$
  - While we have shown that the mean and variance of the product of two independent random variables are $\mu_X\mu_Y$ and $\sigma_X^2\sigma_Y^2+\mu_X^2\sigma_Y^2+\mu_Y^2\sigma_X^2$, respectively, independent of the distribution of the variables, this does not apply to the variance (dependent variables)
  - The distribution of the variance is specific to the distribution of the variable. **Only for some cases are general distributions of the variance easily known and usable**

44

The Chi-Squared Distribution

- The χ_k^2 distribution is a family of distributions of the sum of squares of *k* variates independently drawn from a standard normal distribution
  - For k=1 this corresponds to the distribution of the variance of a standard normal distribution. For larger k it corresponds to the distribution of the variance of a sample set of size k+1 taken from the standard normal distribution. K (sometimes ν) is the degrees of freedom.

$$p_{\chi^2}(x;k) = \frac{x^{k/2-1}e^{-x/2}}{2\Gamma(k/2)}$$

where Γ(n) is the gamma function, a real valued extension of the factorial function whit ist argument decremented by one:

If n is an integer, then: $\Gamma(n) = (n-1)!$

otherwise: $\Gamma(n) = \int_0^\infty e^{-t}x^{n-1}dx$

Chi-Squared Tests

- The χ_k^2 distribution can be used (in a similar fashion to z and t tests) in tests related to the variance of a distribution if the compared to distribution has a known variance.
  - **If the distribution is known to be normal** the χ^2 value for the correct number of degrees of freedom can be used
  - **For distributions of unknown shape**, the distribution of the mean of samples of **sufficient size** is normal with mean μ and standard deviation σ/√n. χ^2 thus models the variance of sample means of sufficiently large samples (after scaling)
  - Example Hypotheses (one sample):

$$H_A: \sigma_X > \sigma_Y \quad \text{thus:} \quad H_0: \sigma_X = \sigma_Y$$

  - Compute χ^2-value:

$$\chi^2_{n-1} = \frac{(N-1)s_X^2}{\sigma_Y^2}$$

  - Then the usual "stuff": translate χ^2-value to the corresponding p-value according to the χ^2 distribution with n-1 DoF, and evaluate significance
    - Translation usually uses χ^2-table. E.g. p = 2.5% -> χ_9^2=19.02

Chi-Square, Really?

- So, really, what can Chi-square be used for?
- It measures the squared difference between a sample and a known random variable.
- Thus, we can check how well a sample fits an assumed distribution. Pearson's Chi-Squared:
  - Goodness of fit: whether or not an observed frequency distribution differs from a theoretical distribution.
  - Test of independence: assesses whether paired observations are independent of each other
- Recommended "recipes" book: J. Crawshaw and J. Chambers, "A Concise Course in Advanced Level Statistics".

47

The F-Distribution

- If the shape of the sample distribution is not known and there are not enough samples to perform a variance test on a normal sampling distribution then we need a different distribution for the variance
- The F-distribution models the distribution of the ratio of two random variables each having a chi-square distribution, with the one in the nominator having a DoF k_1 while the one in the denominator that of k_2, scaled by their perspective DoF

$$\text{I.e., } \frac{X/k_1}{Y/k_2} \sim F(x, k_1, k_2) \text{ if } X \sim \chi^2_{k_1} \text{ and } Y \sim \chi^2_{k_2}$$

$$p_F(x; k_1, k_2) = \sqrt{\frac{(xk_1)^{k_1} * k_2^{k_2}}{(xk_1 + k_2)^{(k_1+k_2)}}} * \frac{\Gamma(^{(k_1+k_2)}/_2)}{x * \left(\Gamma\left(^{k_1}/_2\right) + \Gamma(^{k_2}/_2)\right)}$$

- For k_1=1 and k_2=n-1, the F-distribution models the distribution of the variance of samples drawn from a student-t distribution for sample size n (n-1 degrees of freedom)

48

F-Tests

- There are multiple scenarios under which we can use the F-distribution to evaluate hypotheses about the variance of a random distribution
  - F-Distribuiton F(k_1,k_2) models the ratio of the variances of sample sets drawn from two normal distributions.
    - F-test can evaluate how likely it is to obtain a given value for the ratio between a known normal distribution and a sample set taken from that distribution
    - F-test can evaluate the likelihood to obtain a given value for the ratio between two sample sets taken from a normal distribution.
  - F-Distribution F(1,k) models the variance of a set of samples taken from a t-distribution with k degrees of freedom
- The F-test operates in a similar way as the other tests
  - Example Hypotheses: H_A: $\sigma_X > \sigma_Y$ thus: H_0: $\sigma_X = \sigma_Y$
  - Compute F-value: $F_{N_X-1,N_Y-1} = \frac{s_X^2}{s_Y^2}$
  - Translate F-value to the corresponding p-value (percentile) according to the F-distribution for sample sizes n,m and evaluate significance
    - Translation usually uses F-table. E.g. p = 5% -> $F_{9,9}$=3.18

49

Other Test Distributions

- F-tests and χ^2-tests can also be used in paired scenarios and with and without one of the test distributions being known.
- If the variance does not fit the χ^2 or the F-distribution, there are a large range of related test statistics (and distributions) that can be used
  - Many variations of the F-test
  - Yates corrections
  - Cochran-Mantel-Haenszel chi-square tests
  - McNemar's test
  - Likelihood ratio test
  - Other "ANOVA" (analysis of variance) tests
  - Pearson's r
  - And many moooooreee

50

SUMMARY OF TESTS

51

Summary

- To be able to make statements comparing performance derived from experiments it is necessary to show that the differences are not the result of chance
- Benefits
  - Significance tests are a flexible way to evaluate if a hypothesis about the sampling mean (or some similar statistics) has significant support
  - Significance tests can be applied without complete knowledge of the distributions underlying the problem
- Problems:
  - Significance tests only reject the Null hypothesis
    - No direct proof of the hypothesis
  - Significance tests are difficult when trying to evaluate hypotheses that are not involving the mean

52