STAT206

Andrew Codispoti

December 6, 2015

1 Descriptive and Inferential Statistics Terminology

1. Statistics: Science of conducting studies to collect, organize, summarize, analyze and draw conclusions from data.
2. variable:attribute can be diff values
3. Categorical variables(Gender), Quantitative variables (age weight)
4. Discrete var: values that can be counted
5. continuous var: assume all values between any two specific values, measured
6. random: vals determined by chance
7. Data set:collection of data values
8. descriptive statistics: describe situation in data (collect, organize summarize, present data)
9. inferential statistics: make inference from samples to population
10. population:all indviduals under study
11. sample group selected
12. statistical hyp: claim
13. hypothesis testing: decision making process for evaluating a claim about population from samples

study samples of pop and use inferential stats to make inference about pop

1.1 Freq distribution and graphs

raw data in original form organized in some way (i.e. freq distribution). can then be presented in graphs or charts

freq. dist organizes raw data into classes and frequencies.

range of data is large (lower class limit, upper class limit, class boundaries, class width, class midpoint, open-ended distribution)

1.2 General rules for grouped freq. dist.

5-20 classes, class width should be odd to ensure midpoints are integers, mutually exclusive(non overlapping), continuos, exhaustive, equal in width

1.3 general procedure

1. determine classes
   (a) find highest(H) and lowest(L) values,
   (b) range = H - L
   (c) number(N) of classes desired
   (d) width = R/N (round to whole)
   (e) lowest class limit, convenient value less than or equal to the smallest data value.
   (f) add W to get lowr limit of next class
   (g) class boundaries
2. tally data
3. find freq. of tallies
4. cumulative freq

1.4 Histogram

histogram uses vertical bars to rep. freq of classes

different shapes bell shaped, uniform, j-shaped reverse j-shaped, right and left skewed, bimodal, u-shaped

1.5 Stem and leaf plots

uses part of data value as stem and part as leaf to form groups

more informative because shows values

first arrange data in order, separate into classes then plot

look for peaks and gaps, shape of distribution, variability by looking at spread

can put back to back to compare with other data set

1.6 Descriptive Measures

describe data with numerical measures

characteristic or measure from population is a parameter, from sample data is called a statistic

Measures of central tendency(mean median, mode) **don't round until done**

1.6.1 Average

sum divided by total

$$\mu = \bar{X} = \frac{\sum X}{N}$$

rounding rule: final answer should be rounded to one or more decimal place than original

finding mean from grouped frequency distribution

$$\bar{X} = \frac{\sum f \cdot X_m}{n} \tag{1}$$

f is frequency and X_m are midpoints of class

1.6.2 Median

ordered so called data array

half smaller half larger, position $\frac{n+1}{2}$

when even number of values, b/w two given data values

1.6.3 Mode

value that appears the most frequently

data set can have multiple or no modes

1.6.4 properties of central tendency

1. mean needs all values of data, varies less than media and mode, unique, cannot be computed for open-ended freq. distribution, highly affected by outliers

2. median find middle, whether data falls into upper or lower half of the distribution, can be calc for open data dist. less affected by outlier

3. most typical case req., easier to compute, can be computed for categorical or nominal data, not unique

2 Variation

2.1 Measure of Variation

range is simple measure of variability

Variance is measure of variability that uses all the data points, avg. deviation of values from mean.

population variance σ^2 is avg. of squared deviations of values from population mean μ

$$\sigma^2 = \frac{\sum (X - \mu)^2}{N} \tag{2}$$

where N represents total no. of values in pop

sample variance denoted by s^2 sum of squared deviations of the values from sample mean $\bar{X}$ divided by $(n-1)$

$$s^2 = \frac{\sum (X - \bar{X})^2}{n-1} \tag{3}$$

where n is no. values in sample

divided by n-1 b/c unbiased estimate of σ^2. standard deviation is square root of variance **poulation standard deviation**

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum (X - \bar{\mu})^2}{n-1}} \tag{4}$$

sample standard deviation

$$s = \sqrt{s^2} = \sqrt{\frac{\sum (X - \bar{X})^2}{n-1}}$$
$$s^2 = \frac{\sum X^2 - \frac{(\sum X)^2}{n}}{n-1} \tag{5}$$
$$s^2 = \frac{\sum f \cdot X_m^2 - \frac{(\sum f \dot{X}_m)^2}{n}}{n-1}$$

2.2 Note

1. value of s^2 always +
2. sum of (X - $\bar{X}$) is always 0
3. larger value of s^2 or s, larger variability of data (same with σ)

3 Empirical(Normal) Rule

distro. bell shaped

some % of vals fall within 1 standard deviation of mean (X-nS, X+nS)

4 Chebyshev's Theorem

in an dist regardl of shape, proportion of value that fall within k standard deviations will be at least $1 - \frac{1}{k^2}$where $k > 1$

subtract average from larger value, divide diff by standard deviation to get k, use cheb to get %

5 Measure of Position

standard score / z-score subtract mean from observation and divide result by standard deviation $z = \frac{X - \bar{X}}{s} or \frac{X - \mu}{\sigma}$

number of standard deviations falls above or below mean

5.1 Partition Values

5.1.1 percentile

100 equal groups

5.1.2 decile

10 equal groups

5.1.3 quartiles

4 equal groups

5.2 Percentile Equations

find the percentile of a piece of data

$$Percentile = \frac{numberofdatavaluebelowX + 0.5}{n} * 100\% \tag{6}$$

find the position of a percentil

$$c = \frac{n \cdot p}{100} \tag{7}$$

if c is not a whole number round to next whole number, and position is the number. if c is a whole number average c and c+1. this is position of X

6 Box-Plots

show distrivution of data

helpful for finding outliers in the data based on the min and max relative the Q_1 and Q_3 measurements

modified box plots dont show outliers

7 Product rule

total number of possibilities are $k_1 \cdot k_2 \cdot \ldots k_n$

8 Permutation Rule

permutation: arrangement of n objects in specific order

Rule 1: arrangement of n objects in specific order using r object of time, denoted by nPr:

$$nP_r = \frac{n!}{(n-r)!} \tag{8}$$

arrangment of n objects in specific order taking all n at a time is n!

9 Combination Rule

Combination: selection of objects with no order

the selection of r objects out of n objects can be done in nC_r ways

$$\binom{(}{n}, r) = \frac{n!}{(n-r)!r!} \tag{9}$$

10 Combinatorics Ex

$$n = \frac{8!}{5! * 3!} = \frac{5! * 6 * 7 * 8}{1 * 2 * 3} = \frac{\frac{6!}{3!3!}}{56} + \frac{\frac{6!}{5!}}{56} = \frac{13}{28} \tag{10}$$

11 Bayes Theorem

$$P(A \cap B) = P(A) \cdot P(B|A)$$

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

$$P(A|B) = \frac{P(Ai * P(D|Ai))}{P(Ai)P(D|Ai) + \cdots + P(Ak)P(D|Ak)}$$

12 Probability Distributions

1. variable: characteristic or attribute that can assume different values
2. Random Variable: variable whose value is determined by chance
3. have requirment that $\sum P(X) = 1$, $0 \leq P(X) \leq 1$
4. Expectation: $E(X) = \mu = \sum X \cdot P(X)$

13 The Binomial Distribution

binomial experiment: probability experiment that satisfies:

1. each trial can have only two outcomes or can be reduced to success or failure
2. fixed # of trials
3. outcomes independant of each other

4. probability of success must remain same for each trial

$$Mean : \mu = \sum(X \cdot P(X))$$
$$Variance : \sigma^2 = \sum[X^2 \cdot P(X)] - \mu^2$$
$$StandardDeviation : \sigma = \sqrt{\sigma^2}$$
$$Expectation : E(X) = \mu = \sum(X \cdot P(X))$$

14 Poisson distribution

discrete probability distribution that is useful in large numbe of trials with small success rate

$$P(X; \lambda) = \frac{e^{-\lambda}\lambda^x}{X!} where X = 0, 1, 2 \tag{11}$$

where $\lambda = n \cdot p$ is the parameter of the poisson distribution

1. Mean: $\mu = \lambda$
2. Variance: $\sigma^2 = \lambda$
3. Standard Deviation: $\sigma = \sqrt{\lambda}$

15 Poisson Process

events occur randomly in time and space

1. Independance: number of occurences in disjoint intervals are independant
2. individuality: events occur singly P(two or more events occur simultaneously) = 0
3. Homogeneity: events occur according to a uniform rate of intensity

if events occur with average rate of λ per unit of time, and X is num of events which occur in t units of time, the X Poisson($\lambda \cdot t$)

$$f(x) = \frac{e^{-\lambda t}(\lambda t)^x}{x!}$$
$$E(x) = \lambda t$$
$$Var(x) = \lambda t$$

16 Geometric Distribution

Bernoulli trials completed until successful.

let X be no of independant Bernoulli trials until the first success including first success then X follow geo dist. X Geom(p)

$$f(x) = p(1-p)^{x-1}, x = 1, 2, \ldots$$
$$E(X) = \frac{1}{p}$$
$$Var(X) = \frac{1-p}{p^2}$$

16.1 Memoryless Property of Geometric Random Variable

let X be a Geometric(p) R.V. and t1, t2 $\in$ R. then $P(X = t1 + t2 \mid x = t1) = P(X > t2)$

buy a car and travel x kilometers. reliability does not depend on previous examples

$$LS = \frac{P(X \geq t_1 + t_2 and x \geq t_1)}{P(X \geq t1)} \tag{12}$$

17 Chebyshev's Theorem

Let X be a R.V with E(X) and $Var(x) = \sigma^2$ Then for any $\epsilon > 0$

$$P(|X - E(X)| \geq \epsilon) \leq \frac{Var(x)}{\epsilon^2} \forall \epsilon > 0 \tag{13}$$

$$P(|X - E(X) < \epsilon) = 1 - P(|X - E(X) \geq \epsilon) \geq 1 - \frac{VAR(X)}{\epsilon^2} \tag{14}$$

Let $\epsilon = k\sigma$,where $k \geq 0$, $\sigma > 0 - stdeviation$

$$P(|X - E(X)| < k\sigma) \geq 1 - \frac{\sigma^2}{k^2\sigma^2} \tag{15}$$

$$P(|X - E(X)| < k\sigma) \geq 1 - \frac{1}{k^2} \tag{16}$$

17.1 Example

The # of students who miss a friday class is a Random varible X with mean 15 and st. deviation 2. Find/estimate the probability.

$$P(9 < x < 21) = P(-6 < x - 15 < 6) = P(|X - 15| \leq 6) = 1 - \frac{1}{3^2} \tag{17}$$

17.2 Example 2

$$X\ Bin(5,1/2); FindP(|X-E(X)|<2\sigma)$$
$$E[X]=n.p=5\cdot 1/2=2.5$$
$$Var[X]=np(1-p)=5\cdot 1/21/2=5/4=1.25$$
$$P(|X-2.5|<2\cdot\sqrt{1.25})\geq 1-1/4=0.75$$

actual probability

$$P(|x-2.5|<2\sqrt{1.25}>)=P(-2.236<x-2.5<2.236)$$
$$P(0.264<X<4.37)=P(1\leq x\leq 4)=F(4)-F(0)=0.9688-0.0313=0.9375$$

18 Continuous Random variables and distributions

function from sample space to real numbers

$$def: X: S\rightarrow R[a,b]\in R \tag{18}$$

where R(X) is continuous and individual points in R must have 0 probability

1. $P(X=x)=0 for any x\in\Re$
2. $P(a\leq X\leq b)=P(a<X<b)$

18.1 P

.d.f probabiility density function

for random variable X, denoted f(x), assigns probability to $x\in\Re(X)$

$$P(a<X<b)=\int_a^b f(x)dx,(a,b)\subseteq\Re(x)$$
$$f(x)=0,\forall x\in\Re(x)$$
$$\int_\infty^\infty f(x)dx=1$$

18.2 Cumulative distribution function(cdf)

cdf of continuous random variale X, denoted F(x) gives prob that X takes on value less than or equal to x.

$$F(x) = P(X \leq x) = P(X < x) \tag{19}$$

18.2.1 Properties

1. $F(-\infty) = 0$
2. $F(\infty) = 1$
3. F(x) non decreasing

18.3 Relationship b/w pdf and cdf

$$\begin{aligned}\int_a^b f(x)dx &= P(a < X < b) \\ &= P(X < b) - P(X < a) \\ &= F(b) - F(a)\end{aligned}$$

fundamental theorem of calc $\frac{dF(x)}{dx} = f(x)$

19 Continuous Uniform distribution

the probability of any subinterval of the range is proportional to the length of the interval(two sub-interv. with same length must have same probability)

$$f(X) = \frac{1}{b-a}, a \leq x \leq b$$

$$F(x) = \begin{array}{ll} 0 & x < a \\ \frac{x-a}{b-a} & a \leq x \leq b \\ 1 & b < x \end{array}$$

19.1 Mean/expected value

$$E(X) = \int_x xf(x)dx = \mu$$
$$E[g(x)] = \int_x g(x)f(x)dx$$
$$E[aX + bY] = aE[x] + bE[y]$$

19.2 Variance

$$Var(X) = E[(X - E(X))^2]$$
$$Var[aX1 + bX2] = a^2Var(X1) + b^2Var(X2)$$

19.3 Exponential Distribution

Events occur according to a Poisson process, measure the inter arrival time s b/w events. If X is the amount of time until next event in a posisson process, $X\ Exp(\theta)$ where $\theta = \frac{1}{\lambda}$

$$f(x) = (1/\theta)e^{\frac{x}{\theta}} F(x) = 1 - e^{-\frac{x}{\theta}} E(x) = \theta Var(X) = \theta^2$$

20 Normal Distribution

1. range - to + infinity
2. denote x $X\ N(\mu, \sigma^2)$

$$pdf = f(x) = \frac{1}{\sqrt{2 * \pi\sigma^2}} e^{-1/2(\frac{x-\mu}{\sigma})^2} \tag{20}$$

mean and variance are parameters

1. linear combination of independantly normally distributed parameters is normallydistributed.
2. $P(Z > z) = P(Z < -z)$because of symmetry
3. probabily density function for normally distributed randome variable is not integrable for anyfinite limits a,b.

$$E(X) = \mu$$
$$Var(X) = \sigma^2$$

20.1 Central Limit Theorem

use normal distrubtion to approximate probabilities for non-normal distributions

independant random variables:have no influence on one another's values

$$f(x,y) = P(X = x \cap Y = y) = P(X = x)P(Y = y) = f_x(x)f_y(y) \quad (21)$$

20.1.1 Sum

let $X_a, X_2, \ldots, X_n$ be independant random variables with same distribution $E(X_i) = \mu$ and $Var(X_i) = \sigma^2$

$$asn \to \infty, f\Sigma_i X_i approaches cdf for N(n\mu, n\sigma^2) \quad (22)$$

$$cdf of \frac{\Sigma_i X_i - n\mu}{\sigma\sqrt{n}} \to (0,1) \quad (23)$$

20.1.2 Average

let $X_1, \ldots, X_n$ be independant random variables, with same distribution $E(X_i) = \mu$ and $Var(X_i) = \sigma^2$

as $n \to \infty$ cdf of random variable $X \bar{=} \Sigma_i X_i/n$ approaches cdf $N(\mu, \frac{\sigma^2}{n})$

cdf of randome var $\frac{X \bar{-} \mu}{\frac{\sigma}{\sqrt{n}}}$ approaches cdf for N(0,1)

1. use it when n is large but finite.
2. usually need n $\geq$30
3. works for any variables unless $\mu or \sigma^2$ do not exist

20.1.3 Approximation to Binomial Distribution

$$\frac{X - n * p}{\sqrt{n * p * (1-p)}} \quad (24)$$

20.1.4 Continuity Correction

improve approximation to sum or average of discrete random variables using normal random variable.

go +- 0.5 for bar of width one as the integer value will be centered

21 CI for μ when σ known or n $\geq$30

point estimate of a parameter is a specific numeric value.
interval estimate : interval or range of values used to estimate the parameter

$$\bar{X} - z_{\alpha/2}(\frac{\sigma}{\sqrt{n}}) < \mu < \bar{X} + z_{\alpha/2}(\frac{\sigma}{\sqrt{n}}) \tag{25}$$

point estimate +- maximum error $z_{\alpha/2}(\frac{\sigma}{\sqrt{n}})$ maximum error $z_{\alpha/2}$ critical value α level of significance

21.1 Confidence level

probability that interval estimate will contain the parameter

21.2 confidence interval

specific interval estimate of a parameter determined by using data obtained from sample and by usign specific confidence interval of estimate.

21.3 Sample size

formula for minimum sample size needed for an interval estimate of population mean. E is maximum error

$$n = (\frac{z_{\alpha/2} \cdot \sigma}{E})^2 \tag{26}$$

22 CI for μ when σ known or n ¡ 30

$$\bar{X} - t_{\alpha/2}(\frac{s}{\sqrt{n}}) < \mu < \bar{X} + t_{\alpha/2}(\frac{s}{\sqrt{n}}) \tag{27}$$

$t_{\alpha/2}(\frac{\sigma}{\sqrt{n}})$ maximum error

where s is sample standard deviation and $t_{\alpha/2}$ are found in t-table for n-1 degree of freedom

23 CI and sample size for proportions

1. p = population proportion
2. $\hat{p}$ = sample proportion or $\hat{p} = \frac{X}{n}$ x is # of sample units wiht characteristic n sample size

23.1 sample size for proportions

formula for minumum sample size needed for interval estimate of population is

$$n = \hat{p}\hat{q}(\frac{z_{\alpha/2}}{E}), Eismaximumerrorforproportion \tag{28}$$

24 Maximum likelihood estimator

find derivative of the multipliction of the equatino for all xi's; when multiple unknown parameters, take partial derivatives and solve for unknown.

24.1 Unbiased Estimator

$$E(\tilde{\theta}) = \theta \tag{29}$$

25 Confidence Intervals

$$\bar{X} - \bar{Y} \pm Z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, \sigma = Sforsmall \tag{30}$$

26 Independant small samples

26.1 Pooled Estimator

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} \tag{31}$$

$$dof = n1 + n2 - 2 \tag{32}$$

27 Unequal variace dof

$$\frac{(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2})^2}{\frac{\frac{S_1^2}{n_1}}{n_1 - 1} + \frac{\frac{S_2^2}{n_2})^2}{n_2 - 1}} \tag{33}$$

28 Difference of means–Not Independant

$$d_i = x_i - y_i$$

$$\frac{\bar{D} * \mu_D}{\frac{S_D}{\sqrt{n}}} \; t_{n-1}$$

$$S_D^2 = \frac{\sum_{i=1}^{n}(D_i - \bar{D})^2}{n-1}$$

29 Hypthesis Testing

1. Specify single default hypothesis and check whether the data is unlikely under the hypothesis. Also called null hypothesis because it means a new treatment has no effect.H0
2. Alternative hypothesis. H0 is not true.
3. Test Statistic (discrepancy measure): some funciton fo data that is constructed to measure degree of agreement between the data and null hypothesis.

29.1 Types of Error

TYPE 1 ERROR: reject H0 when it is true.
TYPE 2 ERROR: do not reject H0 when it is false.

$$P(Type1Error) = \alpha(significancelevel)$$
$$P(Type2Error) = 1 - \beta(Pisthepowerofthetest)$$

29.2 Steps

1. Step 1: Determine null and alternative hypothesis.
2. Step 2: Choose a test statistic. $T = T(X_1 \dots X_n; \theta)$.
3. Step 3: Find observed value of T. $T_{obs} = T(x_1, \dots x_n; \theta)$.
4. Step 4: Critical value or p-value. If critical value, find critical value for level of significance, and take into account if its two sided or not. For pvalue determine error amount for the observed value.
5. Step 5: Compare crtical or pvalue with hypothesis and reject if necessary.

29.3 Hypothesis test for mean

$$Z = \frac{\bar{X} - \mu_0}{\frac{\sigma}{n}} \tag{34}$$

29.4 P-Value approach

once find test statistic, find probability greater than.

29.5 Population proportion

$$Z_{obs} = \frac{\hat{p} - p}{\sqrt{\frac{p*(1-p)}{n}}} \tag{35}$$

$$\hat{p} = \frac{X}{n} \tag{36}$$

30 Chi Squared

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1} \tag{37}$$

$$U = \frac{(n-1)S^2}{\sigma^2}, n-1dof \tag{38}$$

$$(\frac{(n-1)s^2}{\chi^2_{\alpha/2}}, \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}} \tag{39}$$

31 Tests about Independance

31.1 Pearson's Test

$$D = \sum_{j=1}^{m} \frac{(O_j - E_j)^2}{E_j} \tag{40}$$

where O is observed and E is expected. For a large n, D has an approximate chi squared distribution.

with r x c table, (a-1)(b-1) degrees of freedom, a num rows, b num of columns

32 Goodness of Fit Test

Try to fit distribution to given data, use pearson's test statistic.

33 Regression

1. Statistically Dependant
2. independant

33.1 Definitions

1. ρ population's correlation coefficient
2. e- sample correlation coefficient

$$\begin{aligned} \rho &= \frac{cov(x,y)}{\sigma_x \sigma_y} \\ cov(x,y) &= E[X - E(X)][Y - E(Y)] \\ cov(x,x) &= E[X - E(X)]^2 \\ -1 \leq p \leq 1 \end{aligned}$$

1. p =¿ +1 strong positive correlation
2. p =¿ -1 strong negative correlation
3. p =¿ 0 weak correlation
4. p == 0 no correlation

$$t_{cal} = r\sqrt{\frac{n-2}{1-r^2}}, n-2 \tag{41}$$

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

$$S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - \frac{(\sum_{i=1}^{n} y_i)^2}{n}$$

$$S_{xy} = \sum_{i=1}^{n}(x_i = \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - \frac{(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n}$$

$$b = \frac{S_{xy}}{S_{xx}}$$

$$\bar{y} = b\bar{x} + a$$

$$s_e^2 = \frac{S_{yy} - (S_{xy})^2/S_{xx}}{n-2}$$

$$t = \frac{b - \beta}{\frac{s_e}{\sqrt{S_{xx}}}}$$

$$b \pm t_{\alpha/2} * s_e/\sqrt{S_{xx}}$$

34 Correlation

$$r = \frac{1}{n-1}\sum_{i=1}^{n}(\frac{x_i - \bar{x}}{s_x})(\frac{y_i - \bar{y}}{s_y})$$

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$