Stat 104: Quantitative Methods

Last Class: Course Review

Important Dates

- Final Exam: December 9, 2pm-5pm
- [extension students-see the email I sent]
- Office hours all next week
- Review Session December 6 11:30am
- Final Exam Discussion Board on Canvas

So I sat there for the entire semester thinking "what a horrible Introduction to Latin Class".

General Course Concepts

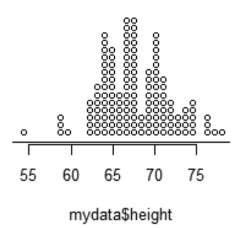
- Visualize
  - Organizing and displaying data (descriptive statistics)
- Conceptualize
  - Methods for data collection (observational studies, sample surveys and controlled experiments)
- Analyze
  - Probability theory and sampling distributions
  - Using samples to make inferences about populations
  - Inference for means and proportions
  - Linear regression

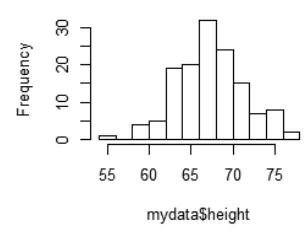
Visualize- Displaying Data

- Population and Samples
- Graphs - boxplots, dotplots & histograms
- Measures of center and spread
- Empirical Rule, Chebysev's Rule

- Relationships between 2 variables
- Scatterplots and correlation (*r*)
- Least-squares regression

Dotplot and Histogram





Summarizing Data

$$s^2 = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{n-1}$$

$$s = \sqrt{\frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{n-1}}$$

```
> library(psych)
> describe(mydata$height)
   vars   n  mean   sd median trimmed  mad min max range  skew kurtosis   se
X1    1 137 67.75 4.07     68   67.68 4.45  54  78    24 -0.03     0.36 0.35
```

n

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

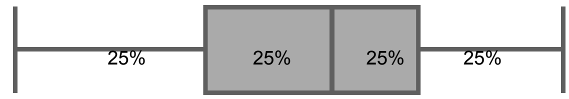
```
> summary(mydata$height)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   54.0    65.0    68.0    67.7    70.0    78.0
```

$Q1$ $Q2$ $Q3$

Box and Whisker Plot

- A Graphical display of data using a 5-number summary:

Minimum -- Q1 -- Median -- Q3 – Maximum
IQR = Q3-Q1
Outlier if point < Q1-1.5IRQ or point > Q3+1.5IQR



Empirical and Chebysev's Rule

- If data is "mound shaped", approximately 95% of the data is in the interval

$$(\overline{x} - 2s_x, \overline{x} + 2s_x) = \overline{x} \pm 2s_x$$

- Without any assumptions, the proportion of the data that lies within k standard deviations of the mean is at least:

$$1 - \frac{1}{k^2}$$

Correlation and Covariance

- Measures of association between two variables.
- Correlation also gives a measure of strength of the relationship (covariance does not).
- These ideas also work for random variables.
- Note that Cov(X,X)=Var(X).

$$r_{xy} = \frac{s_{xy}}{s_x s_y} \qquad s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$$

Correlation and Covariance

```
> cor(cbind(hair,sleep,exercise,height,heartrate),use="complete.obs")
              hair     sleep exercise    height heartrate
hair       1.00000  0.141635  0.15674  0.122391 -0.108267
sleep      0.14164  1.000000  0.19585  0.017132 -0.029111
exercise   0.15674  0.195850  1.00000  0.324190 -0.207153
height     0.12239  0.017132  0.32419  1.000000 -0.036898
heartrate -0.10827 -0.029111 -0.20715 -0.036898  1.000000

> cov(cbind(hair,sleep,exercise,height,heartrate),use="complete.obs")
              hair     sleep exercise    height heartrate
hair       0.95249  0.137531  0.84829  0.484249  -0.91877
sleep      0.13753  0.989918  1.08056  0.069101  -0.25185
exercise   0.84829  1.080565 30.75082  7.288096  -9.98848
height     0.48425  0.069101  7.28810 16.435139  -1.30067
heartrate -0.91877 -0.251851 -9.98848 -1.300673  75.60663
```


Basic Probability Concepts

- Some notation

Idea	Phrase	Concept	Notation
Intersection	A and B	Both A and B	$A \cap B$
Union	A or B	Either A or B or both	$A \cup B$
Complement	Not A	Opposite of A	$\overline{A}$
Conditional	A given B	Given B has occurred, the chance A occurs	$A \mid B$

Basic Probability Rules and Formulas

Rule Name	Definition
Complement Rule	$P(\bar{A}) = 1 - P(A)$
Addition Rule	$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$
Multiplication Rule	$P(A \text{ and } B) = P(A)P(B), \text{ if } A, B \text{ independent}$
Conditional Probability	$P(A \mid B) = P(A \text{ and } B) / P(B)$
Total Probability	$P(B) = P(A \text{ and } B) + P(\bar{A} \text{ and } B)$ $= P(B \mid A)P(A) + P(B \mid \bar{A})P(\bar{A})$
Independence	$P(A \mid B) = P(A)$

The 2x2 Table

- Suppose an applicant for a job has been invited for an interview.
- The chance that
  - He is nervous is $P(N) = 0.7$
  - The interview is successful if he is nervous $P(S \mid N) = 0.2$
  - The interview is successful if he is not nervous $P(S \mid \bar{N}) = 0.9$
- What is the probability the interview is successful ?

	S	$\bar{S}$	
N	(0.2)(0.7)		0.7
$\bar{N}$	(0.9)(0.3)		0.3

Random Variables

- A variable whose numerical values represent the events of a random experiment.
- Can be continuous or discrete
- Have an associated probability distribution which is the possible values of the random variable together with the probabilities corresponding to those values.

Random Variable Formulas

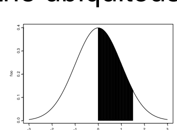
Term	Meaning	Formula
Expected Value μ=E(X)	Long run average	$\mu = \sum x \cdot P(X = x)$
Variance σ^2 =Var(X)	Spread of a random variable	$\sigma_X^2 = \sum_{\text{all } x_i} (x_i - \mu)^2 P(X = x_i)$
Linear Transformation Rule	If X is a rv and Y=a+bX	$E(Y) = a + b\mu_X$ $Var(Y) = b^2\sigma_X^2$
Independence	Knowing X doesn't affect Y	$P_{X \mid Y}(X = x \mid Y = y) = P(X = x)$ $\text{for all } x, y$
Conditional Expectation	Average of X conditional on a y value	$E(X \mid Y = y) = \sum_{\text{all } x \text{ values}} xP(X = x \mid Y = y)$
Mean and Variance of a Sum of Random Variables	$E((a+bX)+(c+dY)) = a + bE(X) + c + dE(Y)$ $Var((a+bX)+(c+dY)) = b^2Var(X) + d^2Var(Y) + 2bdCov(X,Y)$	

The Binomial Random Variable

- A binomial random variable is defined as the number of success in *n* independent trials.
- The binomial random variable is defined by *n*, the number of trials, and *p* the probability of success on any one trial. We write *X~Bin(n,p).*
- The mean of a binomial random variable is *np*.
- The variance of a binomial random variable is *np(1-p).*
- P(*X*=0) = $(1\text{-p})^n$ (all failures)
- P(*X*=n) = $(\text{p})^n$ (all success)
- P(*X*>=1) = 1- $(1\text{-p})^n$ (at least one success)
- P(*X*<n) = 1- $(\text{p})^n$ (at least one failure)

The Normal Distribution

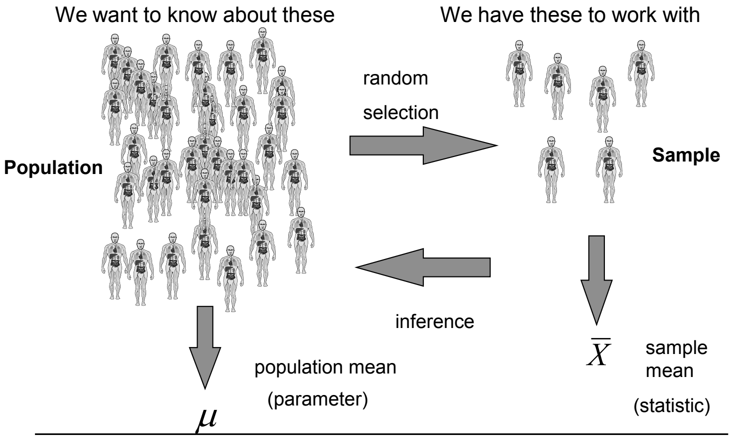
- The normal distribution is the ubiquitous bell-shaped curve.



- We write $X \sim N(\mu, \sigma^2)$
- We usually use the computer to find these probabilities.
- By hand we need to Z-score and use a Z table

$$P(a \le X \le b) = P\left[\left(\frac{a-\mu}{\sigma}\right) \le Z \le \left(\frac{b-\mu}{\sigma}\right)\right]$$

Population versus Sample



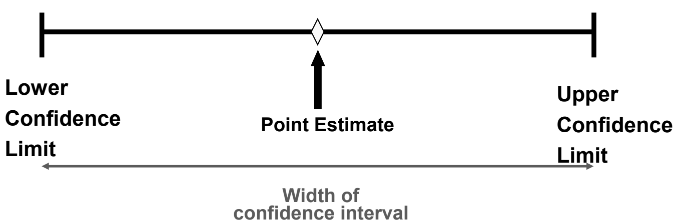
The Central Limit Theorem

- The central limit theorem is one of the more remarkable results in statistics.
- It says that no matter what the underlying population looks like, the distribution of sample means will follow a normal distribution.

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

Point and Interval Estimates

- A point estimate is a single number,
- a confidence interval provides additional information about variability



The Common Point Estimates

We can estimate a Population Parameter …		with a Sample Statistic (a Point Estimate)
Mean	μ	$\bar{x}$
Proportion	p	$\hat{p}$

One Sample Confidence Intervals

- Large sample mean, small sample mean

$$\boxed{\bar{x} \pm z_{\alpha/2} \frac{s}{\sqrt{n}}} \qquad \boxed{\bar{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}}$$

- Large sample proportion, sample size calc

$$\boxed{\hat{p} \pm z_{\alpha/2} \sqrt{\hat{p}\frac{(1-\hat{p})}{n}}} \qquad \boxed{n = \frac{z^2_{\alpha/2}\, p(1-p)}{e^2}}$$

Two Sample Confidence Intervals

- Difference of two means

$$(\bar{X} - \bar{Y}) \pm 1.96 \sqrt{\frac{s_X^2}{n_1} + \frac{s_Y^2}{n_2}}$$

- Different of two proportions

$$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

Remember the *t* Distribution

- The t distribution looks like the N(0,1) distribution except it has **fatter tails**.
- It is centered at zero and defined by its *degrees of freedom* which equal n-1.
- As the sample size n gets large, the t distribution looks like the N(0,1) distribution.

$$t_{n-1} \xrightarrow{n\to\infty} N(0,1)$$

Hypothesis Testing

- Basic approach – set up a null hypothesis H_0 and alternative H_a ; collect data aiming to show H_0 is untrue.
- Two-sided versus one-sided tests
- Reject H_0 if P-value < a priori level (e.g. 0.05) or use test statistic approach.
- P(Type I error) = P(reject H_0 | H_0 is true)

  P(Type II error) = P(not reject H_0 | H_0 is false)

Decision Rules for Testing a Population Mean

$$t_{stat} = \frac{\bar{x}-\mu_0}{s/\sqrt{n}}$$ ⟵ Called the test statistic

$$H_0: \mu = \mu_o \qquad \text{If } |t_{stat}| > 1.96 \quad \textit{reject } H_o$$
$$H_a: \mu \neq \mu_o$$

$$H_0: \mu = \mu_o \qquad \text{If } t_{stat} < -1.64 \quad \textit{reject } H_o$$
$$H_a: \mu < \mu_o$$

$$H_0: \mu = \mu_o \qquad \text{If } t_{stat} > 1.64 \quad \textit{reject } H_o$$
$$H_a: \mu > \mu_o$$

if n <30 use t dist for the cut-off values with df=n-1

Decision Rules for Testing a Proportion

$$T = \frac{(\hat{p}-p_o)}{\sqrt{p_o(1-p_o)/n}}$$

$$H_0: p = p_o \qquad \text{If } |T| > 1.96 \quad \textit{reject } H_o$$
$$H_a: p \neq p_o$$

$$H_0: p = p_o \qquad \text{If } T < -1.64 \quad \textit{reject } H_o$$
$$H_a: p < p_o$$

$$H_0: p = p_o \qquad \text{If } T > 1.64 \quad \textit{reject } H_o$$
$$H_a: p > p_o$$

Two Sample Hypothesis Tests

- Means

  $H_0: \mu_1 = \mu_2$
  $H_a: \mu_1 \neq \mu_2$

  $H_0: \mu_1 = \mu_2$
  $H_a: \mu_1 < \mu_2$

  $H_0: \mu_1 = \mu_2$
  $H_a: \mu_1 > \mu_2$

- Proportions

  $H_0: p_1 = p_2$
  $H_a: p_1 \neq p_2$

  $H_0: p_1 = p_2$
  $H_a: p_1 < p_2$

  $H_0: p_1 = p_2$
  $H_a: p_1 > p_2$

Chi Squares Tests and ANOVA

- Chi Square tests are for hypothesis of the form H_0: $p_1 = a_1$, $p_2 = a_2$, …, $p_k = a_k$
- What is ANOVA used for?

Simple and Multiple Regression

- A single continuous outcome variable, Y, and k predictor variables, $X_1, X_2, \ldots, X_k$
- The statistical model is

  $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon$, $\varepsilon \sim N(0,\sigma^2)$
- Interpretation of least-squares coefficients:

  i) significance, ii) sign, iii) magnitude

  (Change in Y for unit change in X "on average")
- Confidence intervals and prediction intervals
- Assumptions: linearity, constant σ^2, normality of ε

Guide to Regression Output

```
> fit=lm(mydata$hours~mydata$feet)
> summary(fit)

Call:
lm(formula = mydata$hours ~ mydata$feet)

Residuals:
    Min      1Q  Median      3Q     Max
-10.415  -3.429   0.212   3.333  11.908

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.36966    2.07326   -1.14     0.26
mydata$feet  0.05008    0.00303   16.52   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.03 on 34 degrees of freedom
Multiple R-squared:  0.889,     Adjusted R-squared:  0.886
F-statistic:  273 on 1 and 34 DF,  p-value: <2e-16

---
```

$\frac{b_1}{s_{b_1}}$

$s_e = \sqrt{\frac{SSE}{n-k-1}}$

$R^2 = \frac{SSR}{SST}$

Regression Example with Dummy Variables

. regress text_day height male smoke sleep parzen_age

Source	SS	df	MS
Model	61272.5249	5	12254.505
Residual	335947.845	117	2871.3491
Total	397220.37	122	3255.90467

Number of obs =	123
F(5, 117) =	4.27
Prob > F =	0.0013
R-squared =	0.1543
Adj R-squared =	0.1181
Root MSE =	53.585

text_day	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
height	.3089181	1.255918	0.25	0.806	-2.178361	2.796197
male	-13.94764	12.45329	-1.12	0.265	-38.61072	10.71544
smoke	53.69773	22.77152	2.36	0.020	8.599925	98.79554
sleep	-11.50188	4.475374	-2.57	0.011	-20.36512	-2.638637
parzen_age	-2.481976	.9639834	-2.57	0.011	-4.391094	-.5728568
_cons	220.648	100.2586	2.20	0.030	22.09115	419.2048

Regression Diagnostics

- We examine a histogram of the residuals to ensure they look normal. Alternatively we run a normality test on the residuals.
- The standardized residuals are defined as $r_i = e_i/s$. Once the residuals are standardized, they should usually be in the interval (-2,2). If a standardized residual is outside this interval we call it an outlier.
- We plot the standardized residuals versus the fitted values or the x variable. If everything is ok in the regression model we should get a random blob. If we see curvature or extreme points, or funneling out in the regression diagnostic plot that indicates a violation of a regression assumption.

The Final Exam- Topics

More

- Hypothesis tests for the mean, proportion and regression parameters
- Interpretation of regression parameters (and Dummy Variables)
- Assumptions of the regression model
- Confidence intervals for the mean, proportion and regression parameters

- Expectation and variance of random variables (and manipulations)
- Normal distribution, Binomial distribution, Basic probability, ANOVA
- Basic summary statistics (mean, variance, correlation, etc…), Chi Square

Less

Please contact me if there are any questions about exam content/coverage.

Review Time

1) A dummy variable can be assigned up to three values.
   a) True
   b) False

2) Transformations may be used when nonlinear relationships exist between the response and explanatory variable when performing regression.
   a) True
   b) False

3) The value of the coefficient of determination can never decrease when more variables are added to the model.
   a) True
   b) False

Review Time

4) For statistical tests of significance about the regression coefficients, the null hypothesis is that the slope is 1.
   a. True
   b. False

5) If the assumptions of regression have been met, residuals plotted against the independent variable(s) will typically show patterns.
   a) True
   b) False

6) The noise in a regression model is assumed to have zero variance.
   a. True
   b. False

Review Time

11) If the equation of the least squares regression line was computed to be y=45.7+3.1x, then the correlation cannot be less than 0.
   a. True
   b. False

12) If the equation of the regression line that relates percent blood alcohol (x) to reaction time in milliseconds (y) is y=36 - 1.3x, then the slope tells us that for every percent increase in blood alcohol, we can expect reaction time to go down by 1.3 milliseconds
   a. True
   b. False

13) A researcher found the correlation between age of death and number of cigarettes smoked per day to be -0.95. Based just on this information, the researcher can justly conclude that smoking causes early death.
   a. True
   b. False

Review Time

18) A least-squares regression line is not just any line drawn through the points of a scatterplot. What is special about a least-squares regression line?

   a) It passes through all the points.
   b) It minimizes the squared values of the data.
   c) It has slope equal to the correlation between the two variables.
   d) It minimizes the sum of the squared vertical distances of the data points from the line.

Review Time

20) Suppose that the least-squares regression line for predicting y from x is y = 100 + 1.3x. Which of the following is a possible value for the correlation between x and y?

   a) 1.3
   b) –1.3
   c) 0
   d) –0.5
   e) 0.5

Review Time

25) Which of the following is NOT an assumption of the Binomial distribution?

   a) All trials must be identical.
   b) All trials must be independent.
   c) Each trial must be classified as a success or a failure
   d) The number of successes in the trials is counted.
   e) The probability of success is equal to .5 in all trials.

Review Time

34) The weight of a gum drop (piece of candy) in ounces is normally distributed with mean 2 and standard deviation 0.25. A bag contains 10 independent gum drops. The probability that the total weight of the gum drops in the bag exceeds 20 ounces is

   a) 0.25
   b) 0.5
   c) 0.33
   d) 0.75
   e) 0.35

Review Time

36) The purpose of hypothesis testing is to help the researcher reach a conclusion about ______________ by examining the data contained in ______________.

a) a population, a sample
b) an experiment, a computer printout
c) a population, an event
d) a sample, a population

Review Time

37) If the coefficient of determination (R^2) is 0.80, then which of the following is true regarding the slope of the regression line?

a) All we can tell is that it must be positive.
b) It must be 0.80
c) It must be 0.89.
d) Cannot tell the sign or the value.
e) The slope must be significant.

Review Time

39) A multiple regression model with two independent variables exhibits a highly significant F-ratio, but each variable's individual t-statistic is insignificant. The most likely cause of such a situation is

a) Heteroskedasticity
b) Homoskedasticity
c) Multicollinearity
d) Non-normality of residuals

Review Time

41) What is the meaning of the term "heteroscedasticity"?

a) The variance of the errors is not constant
b) The variance of the dependent variable is not constant
c) The errors are not linearly independent of one another
d) The errors have non-zero mean

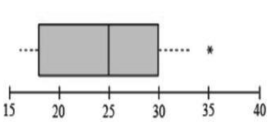
Review Time

61) Suppose we obtain the following regression model for baseball bat sales (Y) when regressed against seasonal indicator variables; $\hat{y} = 100 - 40Spring + 20Wtr - 15Fall$. If we decide to make the baseline season Fall, what would then be the resulting coefficient for Winter (Wtr)?

a) 25
b) -40
c) 30
d) 15
e) None of the above

Review Time

65) Season's Pizza delivers food items to homes in their local area. The following box-and-whisker plot describes the distribution for delivery times in minutes.



Based on this plot, which one of the following statements is correct?

A) The average delivery time is 25 minutes.
B) There are no outliers in this data set.
C) The 75th percentile in this data set is 30 minutes.
D) The second quartile is approximately 18 minutes.
E) None of the above

Review Time

43) Which of the following can NOT be answered from a regression equation?

a) Predict the value of y at a particular value of x.
b) Estimate the slope between y and x.
c) Estimate whether the linear association is positive or negative.
d) Estimate whether the association is linear or non-linear

Review Time

42) Suppose you have estimated wage = 5 + 3education + 2gender – edu*gender, where gender is one for male and zero for female. Suppose instead that gender had been one for female and zero for male. Under this coding what would be the the sum of the coefficients for the gender and interaction variables? (that is we want $b_{gender} + b_{edu*gender}$)

a) -3
b) -1
c) 0
d) 1
e) 2

Finally

- Thanks, and