Probability

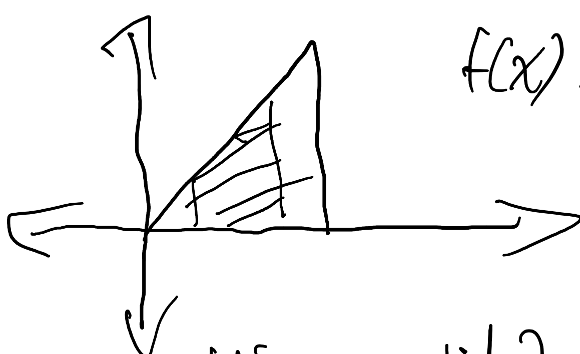
Saturday, August 13, 2016 2:32 PM

- Probability calculus helps understand rules
- Very complicated though, simplify it via **density** and **mass** functions
- Random Variables- numerical outcome of experiment
  - Discrete
  - Continuous
- Probability Mass Function:
  - Discrete
  - Function that associates every outcome with a probability
  - Must always be >= 0
  - Sum of possible vars must always = 1
- Prevalent example:
  - $$(\theta)^x(1-\theta)^{1-x} = p(\theta, x)$$

    θ = prob. of event

    x = ~~#~~ prob. of trials

- Probability Density Function:
  - Continuous
  - Must be larger than 0 everywhere
  - Total area under it must be 1
  - Areas under PDFs correspond to prob for that random var



$$f(x) = \begin{cases} 2x & 0<x<1 \\ 0 & \end{cases}$$

Is this a valid?

Area = 1

Always ≥ 0

Cumulative Distribution Function

Probability var $\leq x$

$$F(x) = P(X \leq x)$$

Survival Function:

Opposite of CDF

$$S(x) = P(X > x)$$

Quantile:

The α^{th} quantile ↔ percentile

CDF → $F(x) = \alpha$

Conditional Probability

Saturday, August 13, 2016 3:15 PM

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

Baye's Rule:

$P(B|A)$ if you know $P(A|B)$

Diagnostic Tests:

$+$ and $-$ events result of diagnostic test

D and D^c event that subject has or does not have disease

retrievable

$P(+|D)$ = sensitivity

$P(-|D^c)$ = specificity

What people care about:

$P(D|+)$

convert $P(D^c|-)$

~~[illegible]~~ Baye's formula:

$P(A|B)P(B)$

$$P(B|A) = \frac{P(A|B)\,P(B)}{P(A|B)\,P(B) + P(A|B^c)\,P(B^c)}$$

Likelihood Ratios:

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

Independence

Saturday, August 13, 2016 3:41 PM

- A is independent of B if:
  - P(A|B) = P(A) where P(B) > 0
  - P(A^B) = P(A)P(B)
- Independent
  - Statistically independent of another event
- Identical Distribution
  - Drawn from the same population distribution

Expected Values

Saturday, August 13, 2016 3:46 PM

- Making conclusions based on noisy data
- Sample expected values will predict population expected values
    - Mean
        - $E[X] = \sum_x xp(x)$
    - Variance
- Population mean is center of mass of population
- Sample mean is center of mass of sample population
- Sample mean is estimate of population mean
- More data in sample mean, closer to actual mean it is

Variability

Saturday, August 13, 2016 4:04 PM

- Variance
  - Variance of random var is measure of spread
  - Var(X) = E[(X-u)]^2
  - Square root of variance is std deviation
  - Sample Variance:

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

Screen clipping taken: 8/13/2016 4:07 PM

- Sample Variance
  - Distribution of sample variance centered at what's being estimated
  - Gets more concentrated and thus closer to population variance with more samples
  - Variance of sample mean is population variance divided by n

Binomial Distribution

Saturday, August 13, 2016 4:31 PM

- Bernoulli Distribution
    - Binomial is a series of Bernoullis
- $(x+a)^n = \sum_{k=0}^{n} \binom{n}{k} x^k a^{n-k}$

Normal Distribution

Saturday, August 13, 2016 4:41 PM

- Centered on mean
- Standard deviations have equal number of values
- 1 std dev away- 68% of data
- 2 std devs- 95%
- 3 std devs- 99%

Poisson Distribution

Saturday, August 13, 2016 4:44 PM

- Used to model counts

$$P(X = x; \lambda) = \frac{\lambda^x e^{-\lambda}}{x!}$$

Screen clipping taken: 8/13/2016 4:45 PM

- Mean and variance are lambda
- Uses:
  - Count data (especially unbounded)
  - Unknown bounds
  - Contingency tables (qualitative values basically quantified)
  - Approximating binomials if n is too large and p is too small
- Poisson random vars represent rates
  - $X \sim Poisson(\lambda t)$
    - $\lambda = E\left[\frac{X}{t}\right]$ is the expected count per unit time
    - t is total monitoring time

Asymptotics and LLN

Sunday, August 14, 2016 12:11 PM

- Asymptotics
    - Behavior of estimators as sample size approaches infinity
    - Frequency interpretation of probabilities
- Law of Large Numbers
    - Average limits what its estimating, the population mean
    - As number of samples goes up, converge to true value
- Estimator is **consistent** if converges to what you want to estimate
- Good estimators must be consitent, with consistent variance and mean

Central Limit Theorem

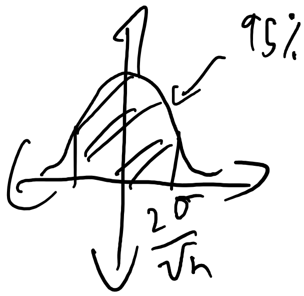
Sunday, August 14, 2016 12:19 PM

- Central Limit Theorem
  - Distribution of averages of iid (Independent Identically Distributed) (properly normalized) approaches standard normal as sample size increases

Confidence Intervals

Sunday, August 14, 2016 12:28 PM

$\bar{X}$ is $\approx$ normal w/ mean μ and sd $\frac{\sigma}{\sqrt{n}}$



$\bar{x} \pm \frac{2\sigma}{\sqrt{n}}$ is 95% confidence interval

For examples, look at Mike's notes

T Confidence Intervals

Monday, September 5, 2016 3:12 PM

T distribution: centered around 0
w/ degrees of freedom as only parameter — not reliant on mean or variance

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \rightarrow \frac{\bar{X}-\mu}{S/\sqrt{n}}$$

T interval is similar to Z-interval, and w/ more data approaches it.

---

t interval

- assumes data are iid normal
- dist. of data symmetric mound shape
- Paired observations
- DON'T use skewed or discrete data, as loses

Hypothesis Testing

Monday, September 5, 2016 3:32 PM

Null hypothesis: H_0 - "status quo"

↳ Attempt to refute

Truth	Decide	Result
H_0	H_0	Correctly accept null
H_0	H_a	Type I Error
H_a	H_0	Correctly reject null
H_a	H_a	Type II Error

Executing a test:

Reject null if $\bar{X}$ > than a constant C

C is chosen to set prob. $\underbrace{\text{Type I}}_{\alpha}$ Error to low, but not too low.

P-values

Monday, September 5, 2016 6:19 PM

- Biggest measure of stat. significance
- Idea: how unusual is it to see current estimate

Approach:

1. Define hyp. distribution
2. Calculate summary/statistic w/ data we have
3. Compare calculation w/ hypothesis, bigger difference = bigger p-value.

Power

Monday, September 5, 2016 6:29 PM

Prob. of rejecting null hypothesis
(when it's false)

β = prob. of type 2 error
(choosing H_a incorrectly)

Power = $1-\beta$

Calculating Power

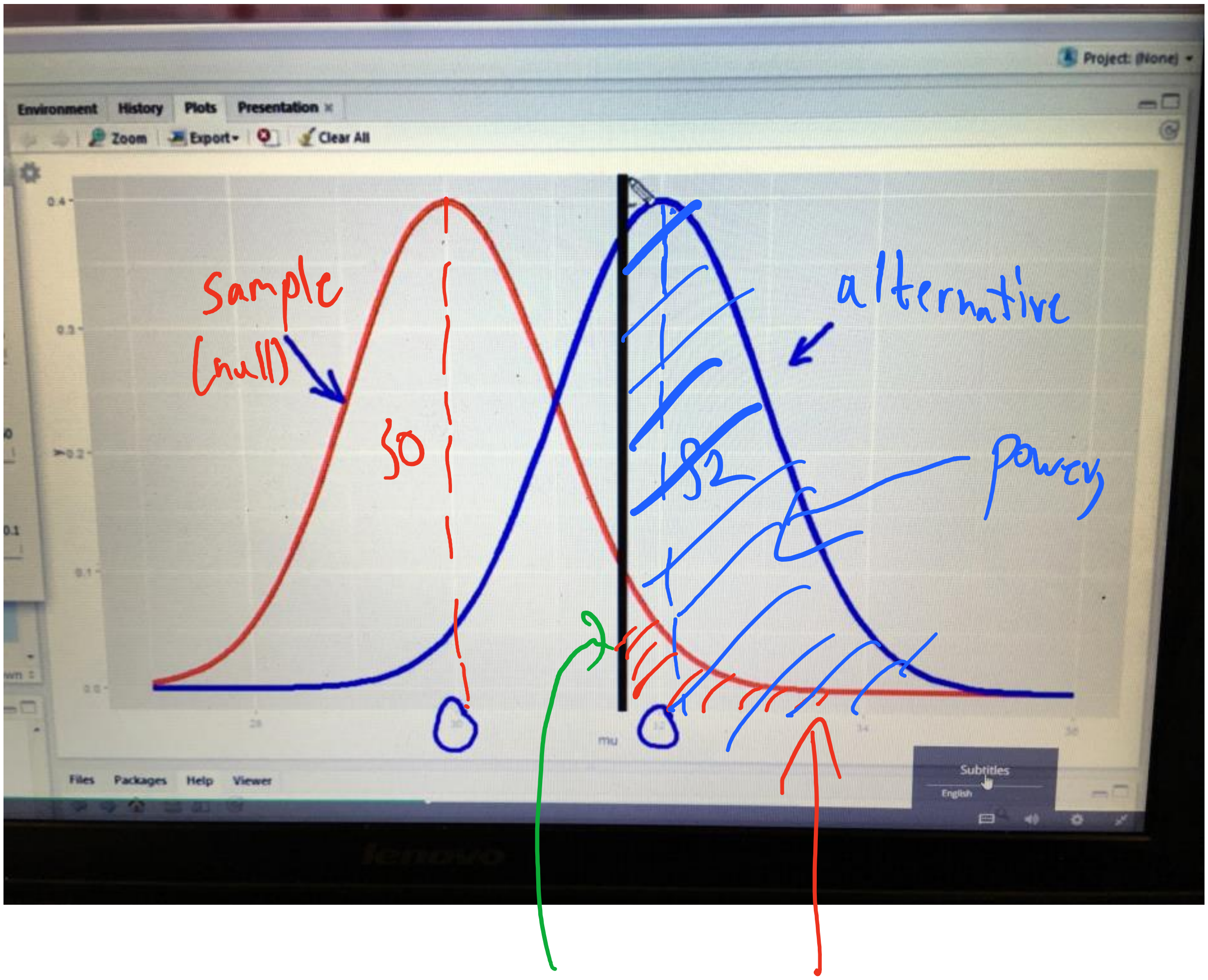
Under H_0: $\bar{X} \sim N(\mu_0, \sigma^2/n)$
Under H_a: $\bar{X} \sim N(\mu_a, \sigma^2/n)$

?? This lecture is hard to understand

$\uparrow$ sample size, $\uparrow$ power

$\uparrow \alpha$, $\uparrow$ power

$\uparrow$ difference μ_a / μ_0, $\uparrow$ power



<5%

Solving Equations in Power

$$1-\beta = P\left(\bar{X} > \mu_0 + Z_{1-\alpha}\frac{\sigma}{\sqrt{n}} ; \mu=\mu_a\right)$$

$$\bar{X} \sim N(\mu_a, \sigma^2/n)$$

Unknown Vars: μ_a, σ, n, β

Vars: μ_a, σ, n, β

Known: μ_0, α

Specify 3 vars, find 4th unknown

T-test Power

Monday, September 5, 2016 7:57 PM

power.t.test — the actual ~~equ~~ application of power, previous page more to explain concept

R function

$$P\left(\frac{\bar{X}-\mu_0}{S/\sqrt{n}} > t_{1-\alpha, n-1}; \mu=\mu_a\right)$$

Multiple Testing

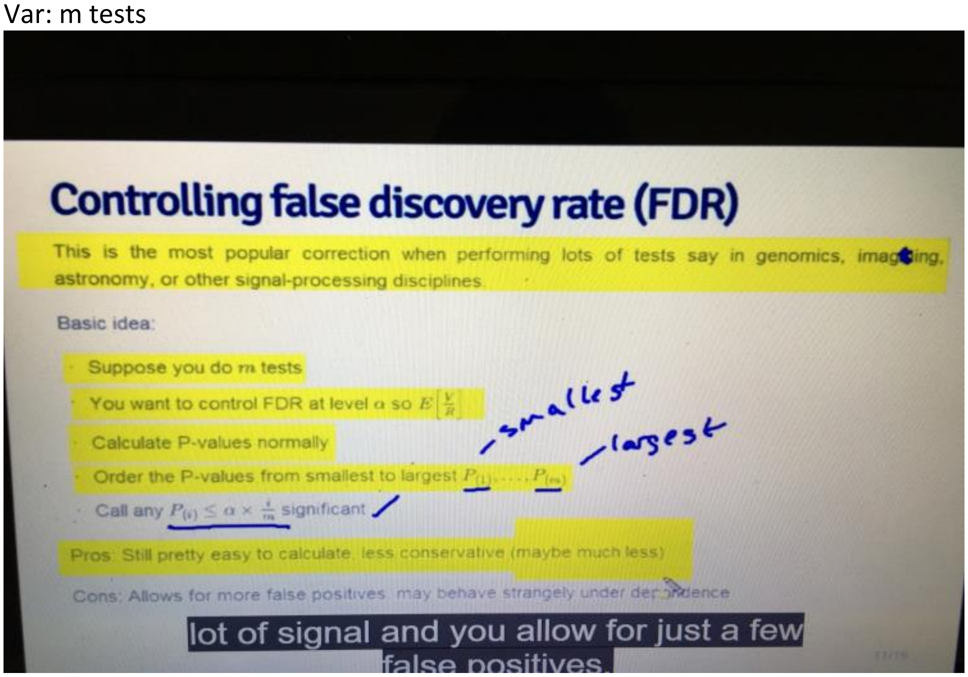
Monday, September 5, 2016 8:59 PM

- Hypothesis testing/significance analysis often overused
- Correcting for multiple testing avoids false positives or discoveries
- 2 components
  - Error measure
  - Correction
- Use large numbers of analyses and data today, small errors pile up
- 2 Types of Errors
  - Type 1- False positives (say there's a correlation, but there's not)
  - Type 2- False negatives (say there's no correlation, but there is)

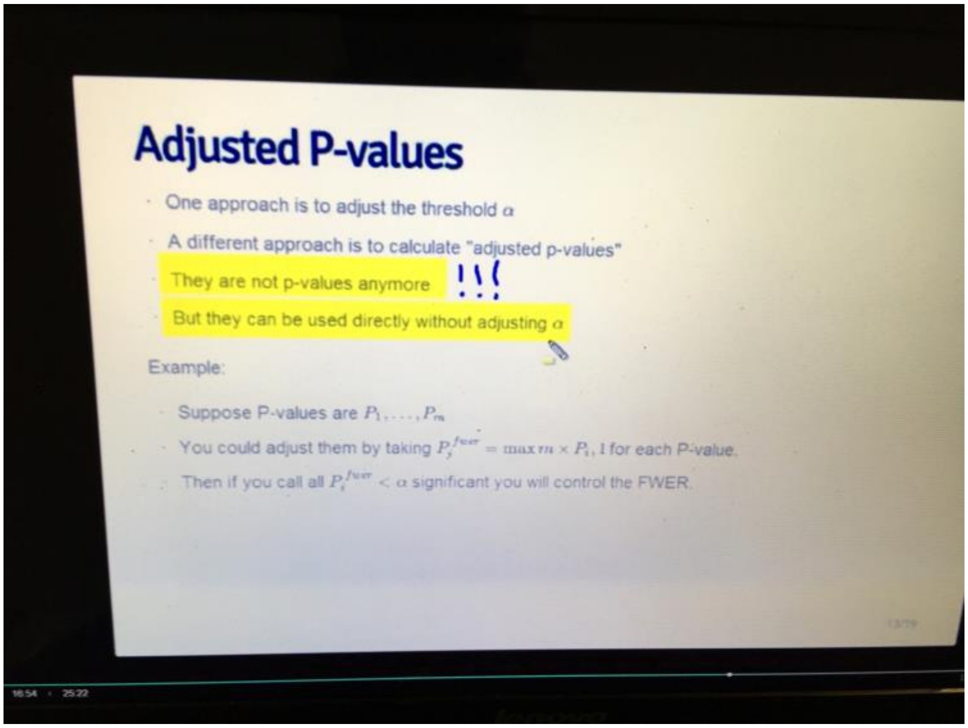
Controlling False Positives

- Bonferroni correction is oldest multiple testing correction
- Approach
  - Var: m = # of tests
  - P-values normally
  - Set a(fwer) = a/m
  - Call all P-values below a(fwer) accurate

Controlling False Discovery Rate (false negatives)

- Var: m tests
- 

Alternative approaches

- 

- ^ Use in conjunction with the other methods

Bootstrapping

Monday, September 5, 2016 9:26 PM

- Bootstrap: useful tool for constructing confidence intervals and calculating standard errors for difficult statistics
- Bootstrap in practice
    - Non-parametric
    - Better percentile bootstrap confidence intervals correct for bias
    - Easily done via R, just feed in data

Permutation Tests

Monday, September 5, 2016 9:39 PM

- Group comparisons
- Null hypothesis that distribution of observations from each group is the same
- Group labels are irrelevant
- Randomly permute labels (and all their data)
- See if permuting the data receives a different result
- Recalculate statistic
    - Mean difference in counts
    - Geometric means
    - T statistic
- Calculate percentage of simulations where the simulated statistic was more extreme