Statistics 67
Introduction to Probability and Statistics for Computer Science

Lecture notes for Statistics

Hal Stern
University of California, Irvine
sternh@uci.edu
From Probability ....

- To this point
  - probability as a measure of uncertainty
  - probabilities for events
    * axioms, probability rules, conditional probability, Bayes’ rule
  - random variables as quantities of interest in an uncertain environment
  - probability distributions as descriptions of possible values for a random variable along with an assessment of how likely each value is to occur
    * discrete/continuous distributions
    * univariate/multivariate distributions
    * joint, marginal, conditional distributions
    * expected values (mean, variance, covariance)
  - sampling/simulation as ways of studying a population or distribution
.... to Statistical Inference

• Goal for remainder of quarter is to use what we know about probability to help us analyze data in scientific studies
  – use a sample from the population to learn about characteristics of the population
  – a common approach is to assume that observed sample are independent observations from a population model (e.g., Poisson or normal)
  – estimate the parameter(s) of the assumed model (e.g., normal mean or binomial proportion)
  – check fit of the assumed probability model
  – draw conclusions based on the estimated parameters (if appropriate)
Point Estimation

• Importance of how data are obtained
  – we don’t discuss in detail here how our data are collected
  – for statistical methods to be valid we need the sample to be representative of the population we are studying
  – typically this involves the use of randomness or chance in selecting the sample to avoid biased selections
  – a simple random sample is the most basic approach and that is what we assume
  – more sophisticated methods (multistage sampling, cluster sampling) can be accommodated
**Point Estimation**

- Estimand – the quantity being estimated

- We can think of two types of estimands
  - Finite population summaries
    * mean of a finite population
    * variance of a finite population
  - Parameters in a mathematical model of a population
    (can think of as an infinite population)
    * $\mu$ or $\sigma^2$ in a Normal distribution
    * $\lambda$ (mean = variance) of Poisson distribution
    * $p$ in a binomial distribution

- For the most part we focus on parameters in a mathematical model of a population
Point Estimation

• Basic Approach
  – suppose \( \theta \) is a parameter that we are interested in learning about from a random sample \( X_1, X_2, \ldots, X_n \)
  – e.g., \( \theta \) might be the mean of the population that we are interested in (\( \mu_X \))
  – \( \hat{\theta} \), a point estimator, is some function of the data that we expect will approximate the true value of \( \theta \)
  – e.g., we might use \( \hat{\mu} = \bar{X} \) to estimate the mean of a population (\( \mu_X \))
  – once we collect data and plug in we have a point estimate \( \bar{x} \)
  – point estimator is the random variable (or function) and point estimate is the specific instance

• Two key questions are
  1. How do we find point estimators?
  2. What makes a good estimator?
Point Estimation - basics

• Assume we have a sample of independent random variables $X_1, X_2, ..., X_n$, each assumed to have density $f(x)$

• We call this a random sample (or iid sample) from $f(x)$

• Assume the density is one of the families we have considered which depends on one or more parameters $\theta$; we usually write the density as $f(x|\theta)$

• Goal is to estimate $\theta$. Why?
  – $f(x|\theta)$ is a description of the population
  – $\theta$ is often an important scientific quantity (e.g., the mean or variance of the population)
**Point Estimation**

Method of moments

- Recall $E(X^j)$ is the $j$th moment of the population (or of the distribution); it is a function of $\theta$
- The $j$th moment of the sample is $\frac{1}{n} \sum_i X_i^j$
- We can equate the sample moment and the population moment to identify an estimator
- Suppose that there are $k$ parameters of interest (usually $k$ is just one or two)
- Set first $k$ sample moments equal to first $k$ population moments to identify estimators
- This is known as the **method of moments** approach
Point Estimation
Method of moments

- Example: Poisson case
  - suppose $X_1, X_2, \ldots, X_n$ are a random sample from the Poisson distribution with parameter $\lambda$
  - recall that $E(X_i) = \lambda$
  - the method of moments estimator is obtained by taking the first sample moment ($\bar{X} = \frac{1}{n} \sum_i X_i$) equal to the first population moment $\lambda$ to yield $\hat{\lambda} = \bar{X}$
  - $Var(X_i)$ is also equal to $\lambda$ so it would also be possible to take the sample variance as an estimate of $\lambda$ (thus method of moments estimates are not unique)
Point estimation

Method of moments

• Example: Normal case
  – suppose $X_1, X_2, \ldots, X_n$ are a random sample from
    the normal distribution with parameters $\mu$ and $\sigma^2$
  – recall that $E(X_i) = \mu$ and $E(X_i^2) = \sigma^2 + \mu^2$
  – to find method of moments estimators we need to
    solve

    \[
    \frac{1}{n} \sum_{i=1}^{n} X_i = \mu
    \]

    \[
    \frac{1}{n} \sum_{i=1}^{n} X_i^2 = \sigma^2 + \mu^2
    \]

  – results:

    \[
    \hat{\mu}_{\text{mom}} = \frac{1}{n} \sum_{i=1}^{n} X_i = \bar{X}
    \]

    \[
    \hat{\sigma}_{\text{mom}}^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2
    \]

• Method of moments summary
  – easy to use
  – generally not the best estimators
  – some ambiguity about which moments to use
Point Estimation

Maximum likelihood estimation

- The density of a single observation is \( f(x|\theta) \)

- The joint density of our random sample is
  \[
  f(X_1, X_2, \ldots, X_n|\theta) = \prod_{i=1}^{n} f(X_i|\theta)
  \]
  (recall the \( X_i \)'s are independent)

- This joint density measures how likely a particular sample is (assuming we know \( \theta \))

- Idea: look at the joint distribution as a function of \( \theta \) and choose the value of \( \theta \) that makes the observed sample as likely as possible

- Likelihood function = \( L(\theta|x_1, \ldots, x_n) = f(x_1, \ldots, x_n|\theta) \)

- Maximum likelihood estimator \( \hat{\theta}_{mle} \) is the value of \( \theta \) that maximizes the likelihood function
Point Estimation
Maximum likelihood estimation

• To find the MLE:
  – solve $dL/d\theta = 0$ to identify stationary point
  – check that we have a maximum
    (can use the 2nd derivative)
  – it is often easier to maximize the logarithm of the likelihood
    (which is equivalent to maximizing the likelihood)
  – in complex models it can be hard to find the maximum
**Point Estimation**

Maximum likelihood estimation

- Example: Poisson case
  - suppose $X_1, X_2, \ldots, X_n$ are a random sample from the Poisson distribution with parameter $\lambda$
  - the joint distribution is
    \[
    f(X_1, \ldots, X_n | \lambda) = \prod_{i=1}^{n} \frac{e^{-\lambda} \lambda^{X_i}}{X_i!}
    \]
  - the likelihood function is
    \[
    L = f(X_1, \ldots, X_n | \lambda) = e^{-n\lambda} \lambda^{\sum_i X_i} / (\prod_i X_i!)
    \]
  - then
    \[
    \log L = \sum_i X_i \ln \lambda - n\lambda - \ln(\prod_i X_i!)
    \]
    \[
    \frac{d\log L}{d\lambda} = \sum_i X_i / \lambda - n = 0
    \]
    which implies that $\hat{\lambda} = \bar{X}$ is the maximum likelihood estimator
  - second derivative of log likelihood confirms this estimate attains a maximum of the likelihood
  - maximum likelihood and method of moments give the same estimator here
Point Estimation

Maximum likelihood estimation

- Example: normal case
  - suppose \( X_1, X_2, \ldots, X_n \) are a random sample from the Normal distribution with mean \( \mu \), variance \( \sigma^2 \)
  - \( \log L = \text{constant} - \frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (X_i - \mu)^2 \)
  - need to solve
    \[
    \frac{\partial \log L}{\partial \mu} = \frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \mu) = 0
    \]
    \[
    \frac{\partial \log L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} (X_i - \mu)^2 = 0
    \]
  - results (same estimators as method of moments)
    \[
    \hat{\mu}_{\text{mle}} = \frac{1}{n} \sum_{i} X_i = \bar{X}
    \]
    \[
    \hat{\sigma}^2_{\text{mle}} = \frac{1}{n} \sum_{i} (X_i - \bar{X})^2
    \]

- Maximum likelihood summary
  - more complex than method of moments
  - statistical theory (not covered) suggests that maximum likelihood estimates do well (especially with lots of data)
**Point Estimation**
Properties of point estimators

- Now have two methods for finding point estimators
- What makes for a good estimator?
  - suppose $T(X_1, \ldots, X_n)$ is an estimator of $\theta$
  - traditional approach to statistics asks how well $T$ would do in repeated samples
  - key to studying estimators is to note that $T$ is itself a random variable and we can study properties of its distribution
  - examples of good properties include
    * lack of bias
    * low variance
Point Estimation
Properties of point estimators

• Unbiasedness
  – estimator $T$ is unbiased for $\theta$ if $E(T) = \theta$
  – unbiased means estimator is ”right on average”
  – no guarantee that the estimate in one sample is good but unbiasedness tells us the estimator does well on average
  – example: in the normal case

$$E(\bar{X}) = \frac{1}{n} \sum_{i} E(X_i) = \mu$$

so $\bar{X}$ is an unbiased estimate for $\mu$

• Variance ($\text{Var } T = E(T - E(T))^2$)
  – suppose we have two unbiased estimators
  – we should prefer the one with low variance
  – but low variance by itself is of limited use - for example $\hat{\theta} = T(X_1, \ldots, X_n) = 6$ (estimator always estimates 6 regardless of the data) has low variance but will be a poor estimate if $\theta$ is far from 6
**Point Estimation**

Properties of point estimators

- **Mean squared error**
  - natural to ask how well $T$ does at estimating $\theta$
  - a difficulty is that we need to know $\theta$ in order to evaluate this
  - $\text{MSE} = E(T - \theta)^2$ is one way to judge how well an estimator performs
  - $\text{MSE}$ depends on $\theta$ but we may find that one estimator is better than another for every possible value of $\theta$
  - it turns out that $\text{MSE} = \text{bias}^2 + \text{variance}$
    (where $\text{bias} = E(T) - \theta$)
  - this yields ... a bias-variance tradeoff
  - consider the example of estimating the normal mean
    * $X_1$ is an unbiased estimator but has a lot of variance
    * $\bar{X}$ is an unbiased estimator but has less variance (dominates $X_1$)
    * $T = 6$ (a crazy estimator that always answers 6!!) has zero variance but lots of bias for some values of $\theta$
Point Estimation
Properties of point estimators

• Large sample properties
  – natural to ask how well \( T \) does in large samples
  – consistency – estimate tends to the correct value in large samples
  – efficiency – estimate has smallest possible variance of any estimator in large samples
  – turns out that maximum likelihood estimators have these good large sample properties
There is one alternative approach to point estimation that we introduce.

It differs from everything else we’ve done in that it allows us to use information from other sources.

Related to Bayes Theorem so known as Bayesian estimation.

Motivation
- suppose we want to predict tomorrow’s temperature
- a natural estimate is average of recent days temperatures (this is like using $\bar{X}$)
- we have other knowledge (typical Southern California weather at this time of year)
- natural to wonder if an estimator that combines information from history with current data will do better.
Point Estimation
Bayesian estimation

• Three components to Bayesian point estimation
  1. Prior distribution $g(\theta)$ describing uncertainty about $\theta$ before any data is examined
  2. Likelihood / data distribution $f(X_1, \ldots, X_n | \theta)$ summarizing the information in the data about $\theta$ (assuming we have the right distribution)
  3. Posterior distribution $p(\theta | X_1, \ldots, X_n)$ is obtained by using Bayes Theorem to combine the prior distribution and the likelihood as
     
     $p(\theta | X_1, \ldots, X_n) = \frac{f(X_1, \ldots, X_n | \theta) g(\theta)}{f(X_1, \ldots, X_n)}$. This posterior distribution describes the uncertainty about $\theta$ after combining the information in the prior distribution and in the data

* A final step is to define an estimator that summarizes the posterior distribution – most common to use the mean of the posterior distribution of $\theta$ as the estimator
Point Estimation

Bayesian estimation

- Bernoulli trials example
  - assume $X_1, \ldots, X_n$ are indep Bernoulli trials with probability of success $\pi$
  - prior distribution for $\pi$ is a uniform distribution between 0 and 1 (completely unsure about $\pi$) so that $g(\pi) = 1$ for $0 < \pi < 1$
  - likelihood is $L = \pi \sum_i X_i (1 - \pi)^{n - \sum_i X_i}$
  - turns out that the posterior distribution is a known continuous distribution (the Beta distribution with parameters $\sum_i X_i + 1$ and $n - \sum_i X_i + 1$)
  - posterior mean (Bayesian point estimator) is $\hat{\pi} = \frac{\sum_i X_i + 1}{n + 2}$
  - note that this is different than $\hat{\pi} = \bar{X}$ which would be the maximum likelihood estimator or the method of moments estimator
  - an interesting case: consider $X = 0$ for which maximum likelihood estimate is $\hat{\pi} = 0$ and for which Bayes estimate is $\hat{\pi} = 1/(n + 2)$
Interval Estimation

• Point estimation is an important first step in a statistical problem

• A key contribution of the field of statistics though is to supplement the point estimate with a measure of accuracy (e.g., the standard deviation of the estimator is such a measure)

• A common way to convey the estimate and the accuracy is through an interval estimate

• In other words we create an interval (based on the sample) which is likely to contain the true but unknown parameter value

• This interval is usually called a confidence interval (CI)

• There are a number of ways to create confidence intervals, we focus on a simple approach appropriate for large samples to illustrate the approach
Central Limit Theorem

- A key mathematical result that enables interval estimation (and other forms of statistical inference) is the central limit theorem (CLT).

- **Theorem:** Let $X_1, \ldots, X_n$ be a random sample of size $n$ from a distribution with mean $\mu$ and variance $\sigma^2$. Then for large $n$, $\bar{X} = \frac{1}{n} \sum_i X_i$ is approximately normal with mean $\mu$ and variance $\sigma^2/n$.

- Note this means $(\bar{X} - \mu)/(\sigma/\sqrt{n})$ is approximately standard normal.

- How big does $n$ have to be? It depends on the population distribution:
  - if the population distribution is itself normal, then the CLT holds for small samples (even $n = 1$)
  - if the population distribution is not too unusual, then the CLT holds for samples of 30 or more
  - if the population distribution is unusual (e.g., very long tails), then the CLT may require 100 or more observations.
Central Limit Theorem - example

- Example: The number of files stored in the home directory has mean $\mu = 7$ and standard deviation $\sigma = 5$. (Note that this variable can not have a normal distribution because: (1) it is a discrete random variable; and (2) with that mean and s.d. the normal distribution would have substantial probability below zero.) What is the probability that a class of 50 students will store more than 400 files?

- First, we should note that the question about the total number of files is equivalent to asking for the probability that $\bar{X}$ will be greater than 8.

- Then by CLT $\bar{X}$ is approximately normal with mean 7 and s.d. $\frac{5}{\sqrt{50}} = .707$

- Finally
  \[
P(\bar{X} > 8) = P\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} > \frac{8 - 7}{.707}\right) = P(Z > 1.41) = .0793
  \]
  (where $Z$ is standard normal random variable)
Central Limit Theorem - binomial proportion

- You may recall that we saw a result something like the CLT in talking about the normal approximation to the binomial distribution – if $np > 5$ and $n(1 - p) > 5$ then $X \sim \text{Bin}(n, p)$ can be approximated by a normal random variable $Y$ having mean $np$ and variance $np(1 - p)$

- This is equivalent to the CLT if we look at the proportion of successes $X/n$ rather than the count of successes $X$

- To be specific, let $W_1, \ldots, W_n$ be a random sample of Bernoulli trials ($0/1$ random variables) with probability of success $p$ (hence mean is $p$ and variance is $p(1 - p)$) and let $X = \sum_i W_i$ be the total number of successes in $n$ trials. Then by the CLT $\bar{W} = X/n$ is approximately normal with mean $p$ and variance $p(1 - p)/n$
Central Limit Theorem - binomial proportion

• Example: Consider sampling light bulbs from a company which claims to produce only 2% defective light bulbs. What is the probability that a sample of 500 light bulbs would yield a defective proportion below 1%?

  – Let $\bar{W}$ equal proportion of defectives in a sample of 500 light bulbs from a population with 2% defectives

  – By CLT $\bar{W}$ is approximately normal (note that $np = 10$ and $n(1 - p) = 490$) with mean $.02$ and variance $(.02)(.98)/500 = .0000392$

  – $P(\bar{W} < .01) = P\left(\frac{\bar{W} - p}{\sqrt{p(1-p)/n}} < \frac{.01 - .02}{\sqrt{.0000392}}\right) = P(Z < -1.60) = .0548$
Interval Estimation
Population mean

• Central Limit Theorem enables us to easily build a confidence interval for the mean of a population

• Assume \( X_1, \ldots, X_n \) are independent random variables with mean \( \mu \) and variance \( \sigma^2 \)

• Then \( \bar{X} = \frac{1}{n} \sum_i X_i \) (the sample mean) is the natural estimate of \( \mu \) (MLE, method of moments)

• We also know that \( \bar{X} \) is a random variable which has approximately a normal distribution, \( \bar{X} \sim N(\mu, \sigma^2/n) \)

• It follows that \( \Pr(-1.96 < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < 1.96) \approx .95 \)

• Thus \( \bar{X} \pm 1.96\sigma/\sqrt{n} \) is an (approximate) 95% confidence interval for \( \mu \)

• Note the above is an exact confidence interval if the population distribution of the \( X \)’s is normal and an approximate confidence interval valid for large \( n \) if not
Interval Estimation
Population mean (cont’d)

• $\bar{X} \pm 1.96\sigma/\sqrt{n}$ is an (approximate) 95% confidence interval for $\mu$ (based on CLT)

• Some variations/improvements
  – Different confidence level
    * We can get a different confidence level by using a suitable percentile of the standard normal distribution
    * e.g., $\bar{X} \pm 1.645\sigma/\sqrt{n}$ is an (approximate) 90% confidence interval for $\mu$
  – Unknown population standard deviation
    * Results given so far require knowing the population standard deviation $\sigma$
    * If $\sigma$ is not known (it usually isn’t) then we can use the sample standard deviation
      $s = \sqrt{\frac{1}{n-1} \sum (X_i - \bar{X})^2}$ as an estimate
    * Then $\bar{X} \pm 1.96s/\sqrt{n}$ is an approximate 95% confidence interval that should be good in large samples (now even larger than before .. say 100 observations or more)
    * It turns out that it is possible to create a more exact 95% confidence interval in this case by replacing 1.96 with the relevant percentile of Student’s $t$-distribution (not covered in this class)
Interval Estimation
Binomial proportion

• Assume $X_1, \ldots, X_n$ are independent Bernoulli trials with probability of success $\pi$ (change from $p$ now)

• Then $\hat{\pi} = \frac{1}{n} \sum_i X_i$ = the sample proportion of successes is the natural estimate (MLE, method of moments)

• From central limit theorem we know that $\hat{\pi}$ is approximately normal with mean $\pi$ and s.d. $\sqrt{\pi(1 - \pi)/n}$

• It follows that $\Pr(-1.96 < \frac{\hat{\pi} - \pi}{\sqrt{\pi(1 - \pi)/n}} < 1.96) = .95$

• Thus any $\pi$ for which the inequality above is satisfied is in a 95% confidence interval

• An alternative is replace the s.d. of $\hat{\pi}$ by an estimate, $\sqrt{\hat{\pi}(1 - \hat{\pi})/n}$ and then note that $\hat{\pi} \pm 1.96 \sqrt{\hat{\pi}(1 - \hat{\pi})/n}$ is an approximate 95% confidence interval for $\pi$
Interval Estimation

- General approach
  - the previous two examples suggest a general approach
  - suppose that we have a point estimator \( \hat{\theta} \) for a parameter \( \theta \)
  - \( \hat{\theta} \) is a random variable with expected value typically approximately equal to \( \theta \) and with a standard deviation \( s.d.(\hat{\theta}) \)
  - it follows that an approximate large-sample 95% confidence interval for \( \theta \) is given by \( \hat{\theta} \pm 1.96 \, s.d.(\hat{\theta}) \) (sometimes we may need to estimate the s.d.)

- Interpretation
  - it is important to remember the interpretation of these confidence intervals
  - the “confidence” belongs to the procedure; we have a procedure that creates intervals having the property that 95% of the confidence intervals contain the true values
  - for any given instance the CI either contains the true value or not; our guarantee is only for average behavior in repeated samples
Tests/Decisions

• Point estimates and interval estimates are important components of statistical inference

• Sometimes there is a desire however for a formal test or decision based on the value of a particular parameter

• For example:
  – We may want to assess whether \( \pi = 0.5 \) in a binomial situation (or in other words we may want to ask if we have a fair coin)?
  – We may want to test whether \( \mu = 0 \) (no change due to an intervention)?
  – We may want to compare average response in two groups to see if they are equal \( (\mu_1 = \mu_2) \)?
Statistical Tests - binomial case

• We illustrate the basic approach in the binomial setting

• Assume we sample \( n \) people at random from list of CS faculty in the U.S.

• Ask each whether their laptop runs Windows or Linux

• Observe 56% use Linux

• Can we conclude that a majority of CS faculty in the US prefer Linux for their laptop?
  – seems obvious that we can but ...
  – the difference between 56% and 50% may just be a fluke of the sample, the truth may be that the population is split 50/50
Statistical Tests - binomial case

• The logic of statistical tests
  – let $X$ denote the number of faculty preferring Linux
  – assume $X \sim Bin(n, \pi)$
    (note we use $\pi$ instead of the usual $p$ to avoid confusion later)
  – organize test in terms of null hypothesis (no effect, no difference) and alternative hypothesis (the difference we suspect may be present)
    * null $H_o : \pi = 0.50$
    * alternative $H_a : \pi > 0.50$
    * why use this formulation? easier to disprove things statistically than to prove them
  – we suspect $H_o$ is false (and $H_a$ is true) if $X/n = \hat{\pi}$ is greater than 0.5. How much greater does it have to be?
  – approach: assume the null hypothesis is true and ask whether the observed data is as expected or is unusual
Statistical Tests - general comments

- There are two slightly different (but related) approaches
  - **significance tests** – assess the evidence against $H_0$
    with a $p$-value that measures how unusual the observed data are
  - **hypothesis tests** – formally establish a decision rule for deciding between $H_0$ and $H_a$ to achieve desired goals (e.g., decide $H_a$ is true if $\hat{\pi} > c$ where $c$ is chosen to control the probability of an error)
  - we focus on significance tests in this class
Statistical Tests - general comments

• The key concept in significance tests is the $p$-value

• $p$-value = probability of observing data as or more extreme than the data we obtained if $H_o$ is true

• Low $p$-values are evidence that either
  (1) $H_o$ is true and we saw an unusual event
  or
  (2) $H_o$ is not true

• The lower the $p$-value the more likely we are to conclude that $H_o$ is not true

• Often use $p < .05$ as serious evidence against $H_o$ but a strict cutoff is a BAD IDEA

• A couple of important points
  – the $p$-value DOES NOT measure the probability that $H_o$ is true
  – even if $p$-value is small the observed failure of $H_o$ may not be practically important
Statistical Tests - binomial case

• Now return to binomial case and suppose that we have sampled 100 professors and find 56 use Linux, or in other words \( n = 100 \) and \( \hat{\pi} = .56 \)

• There are actually two ways to find the \( p \)-value: use the binomial distribution directly or, if \( n \) is large (as it is here) then we can use the CLT

• By the binomial distn ... let \( X \) be number of Linux supporters. Then under \( H_o \) we know \( X \sim \text{Bin}(100, .5) \) and \( P(X \geq 56) = .136 \) (not in our table but can be computed)

• By the CLT ...

\[
p\text{-value} = \Pr(\hat{\pi} \geq 0.56 \mid \pi = 0.5) = \Pr\left(\frac{\hat{\pi} - 0.50}{\sqrt{.5(.5)/100}} \geq \frac{.56 - .50}{\sqrt{.5(.5)/100}}\right) \\
\approx \Pr(Z \geq 1.2) = .115
\]

(using the continuity correction we’d say \( p = P(\hat{\pi} \geq .555) = .136 \))

• Conclude: observed proportion .56 is higher than expected but could have happened by chance so can’t conclude that there is a significant preference for Linux
Statistical Tests - binomial case

- Interpreting results
  - The $p$-value of .136 does not mean that $H_0$ is true, it only means the current evidence is not strong enough to make us give it up
  - $p$-value depends a lot on sample size
    ... with $n = 200$ and $\hat{\pi} = .56$ we would have $p = .045$
    ... with $n = 400$ and $\hat{\pi} = .56$ we would have $p = .008$
Hypothesis Tests

• Significance tests focus on $H_0$ and try to judge its appropriateness

• Hypothesis tests treat the two hypotheses more evenly and are thus used in more formal decision settings
  – hypothesis testing procedures trade off two types of errors
  – type I error = reject $H_0$ if it is true
  – type II error = accept $H_0$ if it is false
  – we can vary cutoff of test; if we increase cutoff to make it harder to reject $H_0$ then we reduce type I errors but make more type II errors (and vice versa if we lower the cutoff)

• In practice hypothesis tests are very closely related to significance tests
Relationship of tests to other procedures

• Tests and confidence intervals
  – confidence intervals provide a range of plausible values for a parameter
  – tests ask whether a specific parameter value seems plausible
  – these ideas are related ... suppose we have a 95% confidence interval for \( \pi \)
    * if \( \pi = 0.50 \) is not in the confidence interval then our test will tend to reject the hypothesis that \( \pi = 0.50 \)

• Tests and Bayesian inference
  – we have not emphasized the Bayesian approach to testing but there is one
  – to see how it might work, recall that the Bayesian approach yields a posterior distribution telling us, for example, the plausible values of \( \pi \) and how likely each is
  – the Bayesian posterior distribution can compute things like \( P(\pi > 0.5|\text{observed data}) \) which seems to directly address what we want to know
Decisions/Tests – general approach

- General setting: we have a hypothesis about a parameter \( \theta \), say \( H_o : \theta = \theta_o \) (could be \( \pi \) in binomial or \( \mu \) in normal) and want to evaluate this null hypothesis against a suspected alternative \( H_a : \theta > \theta_o \)

- A general approach:
  - obtain a suitable point estimate \( \hat{\theta} \) and use it to test the hypothesis (reject \( H_o \) if \( \hat{\theta} \) is far from \( \theta_o \))
  - calculate \( p \)-value which is \( P(\hat{\theta} > \text{observed value}) \) assuming \( H_o \) is true
  - this calculation requires distribution of \( \hat{\theta} \)
  - distribution of \( \hat{\theta} \) will depend on specific example (e.g., binomial case above)

- Of course if alternative is \( \theta < \theta_o \) then \( p \)-value also uses “<”
Decisions/Test – population mean

Example: Tests for $\mu$ (the population mean)

• Natural estimate is $\bar{X}$ (the sample mean)

• What do we know about the distribution of $\bar{X}$ under $H_0$?
  
  – If the population data are normal and $\sigma$ is known, then $\bar{X}$ is normal with mean $\mu_o$ and s.d. $\sigma/\sqrt{n}$
  
  – If the population data are normal and $\sigma$ is not known, then $\bar{X}$ is approximately normal with mean $\mu_o$ and s.d. $s/\sqrt{n}$ for large sample size
  
  – If sample size is large (no matter what the population data are), then $\bar{X}$ is approximately normal with mean $\mu_o$ and s.d. $s/\sqrt{n}$
  
  – Only difference between the last two items is that might expect to need a “larger” sample size in the last case

• The above discussion leads to normal test of $H_o : \mu = \mu_o$ with

\[
p\text{-value} = P(\bar{X} > \bar{x}) = P(Z > (\bar{x} - \mu_o)/\frac{s}{\sqrt{n}})
\]

(with $Z$ the usual standard normal distn)
Decisions/Test – population mean

Example: Tests for \( \mu \) (the population mean)

Some technical stuff (optional)

- When we don’t know \( \sigma \) and plug in the estimate \( s \), we should really adjust for this in our procedure

- It turns out that the proper adjustment (original discovered by a brewery worker!) is to use Student’s \( t \)-distribution in place of the standard normal distribution

- Student’s \( t \)-distribution is a distribution that looks something like the normal but has heavier tails (bigger values are possible). The \( t \) distribution is described by the number of degrees of freedom (how big a sample it is based on) with a large degrees of freedom corresponding more closely to a normal distribution

- Student’s \( t \)-test of \( H_0 : \mu = \mu_o \) would lead to

\[
p-value = P(\bar{X} > \bar{x}) = P(Z > (\bar{x} - \mu_o)/s) / \sqrt{n}
\]

where \( t_{n-1} \) is a random variable having Student’s \( t \)-distribution with \( n - 1 \) degrees of freedom

- For Stat 67 purposes ... just need to know that in large samples can use normal table and not worry about the Student’s \( t \)-distribution
Decisions/Test – population mean

• Numerical example:

Suppose that the average database query response time is supposed to be 1 second or faster. We try 100 queries and observe an average response time of 1.05 seconds (with a standard deviation of .25 seconds). Can we conclude that the database does not meet its standard?

– frame question as a statistical test:
  \[ H_0 : \mu = 1 \text{ vs } H_a : \mu > 1 \]

– \( p \)-value
  \[ = P(Z \geq (1.05 - 1.00)/\frac{0.25}{\sqrt{100}}) = P(Z \geq 2) = 0.023 \]
  (if we use Student’s t-test, then \( p \)-value = .024)

– reject \( H_0 \) and conclude that the database is not performing as advertised

– note that the additional .05 seconds may not be practically important
Decisions/Test – difference between two means

Note to self: If there is time, then do this slide and the next to show how testing handles harder problems

- A common situation is that we have two populations and we want to compare the means of the two populations
- Example (medical): suppose we have two treatments (drug A and drug B) and wish to compare average survival time of cancer patients given drug A ($\mu_1$) to average survival time of cancer patients given drug B ($\mu_2$)
- Assuming we have data on the two populations
  - $\bar{Y}_1 - \bar{Y}_2$ is an estimator for $\mu_1 - \mu_2$
  - $\bar{y}_1 - \bar{y}_2$ is an estimate for $\mu_1 - \mu_2$
  - $\text{Var}(\bar{Y}_1 - \bar{Y}_2) = \sigma^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)$
  - $S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1 + n_2 - 2}$ is a pooled estimator for common variance $\sigma^2$
- Key result: under assumptions
  $$t = \frac{\bar{Y}_1 - \bar{Y}_2 - (\mu_1 - \mu_2)}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$
- Again for Stat 67 don’t worry about Student’s $t$ (for large samples can use normal distribution)
Decisions/Test – difference between two means

- **Confidence interval**
  - 95% confidence interval for $\mu_1 - \mu_2$ assuming large samples is
  \[
  \bar{Y}_1 - \bar{Y}_2 \pm 1.96S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}
  \]

- **Tests of hypotheses**
  - null hypothesis $H_o : \mu_1 = \mu_2$ (no difference)
  - alternative hypothesis $H_a : \mu_1 \neq \mu_2$ (two-sided)
    - or $\mu_1 > \mu_2$ or $\mu_1 < \mu_2$ (one-sided)
  - test statistic $t = (\bar{Y}_1 - \bar{Y}_2) / \left( S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right)$
  - $p$-value = probability of obtaining a value of the test statistic as big or bigger than the observed value if $H_o$ is true (need to use $t$-distribution or normal table if samples are large to find $p$-value)
Probability and Statistical Modeling

• So far:
  – Estimation
    * sample from population with assumed distribution
    * inference for mean or variance or other parameter
    * point or interval estimates
  – Decisions / Tests
    * judge whether data are consistent with assumed population
    * judge whether two populations have equal means
  – To apply statistical thinking in more complex settings (e.g., machine learning)
    * build a probability model relating observable data to underlying model parameters
    * use statistical methods to estimate parameters and judge fit of model
Simple linear regression

Introduction

• We use linear regression as a (relatively simple) example of statistical modeling

• Linear regression refers to a particular approach for studying the relationship of two or more quantitative variables

• Examples:
  – predict salary from education, years of experience, age
  – find effect of lead exposure on school performance

• Useful to distinguish between a functional or mathematical model
  \[ Y = g(X) \quad \text{(deterministic)} \]
  and a structural or statistical model
  \[ Y = g(X) + \text{error} \quad \text{(stochastic)} \]
Simple linear regression
Linear regression model

• The basic linear regression model

\[ Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, \ldots, n \]

- \( Y_i \) is the response or dependent variable
- \( x_i \) is the predictor, explanatory variable, independent variable
- \( x_i \) is treated as a fixed quantity
  (i.e., is not a random variable)
- \( \epsilon_i \) is the error term or individual variation
- \( \epsilon_i \) are independent \( N(0, \sigma^2) \) random variables

• Key assumptions

- linear relationship between \( Y \) and \( x \)
- independent (uncorrelated) errors
- constant variance errors
- normally distributed errors
Simple linear regression

Interpreting the model

- Model can also be written as

\[ Y_i \mid X_i = x_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2) \]

- mean of \( Y \) given \( X = x \) is \( \beta_0 + \beta_1 x \) (known as the conditional mean)
- \( \beta_0 \) is conditional mean when \( x = 0 \)
- \( \beta_1 \) is the slope, measuring the change in the mean of \( Y \) for a 1 unit change in \( x \)
- \( \sigma^2 \) measures variation of responses about the mean
Simple linear regression
Where does this model come from?

• This model may be plausible based on a physical or other argument

• The model may just be a convenient approximation

• One special case is worth mentioning:
  It turns out that if we believe that two random variables $X$ and $Y$ have a bivariate normal distribution (remember we saw this briefly), then the conditional distribution of $Y$ given $X$ is in fact a normal model with mean equal to a linear function of $X$ and constant variance
Simple linear regression

Estimation

• Maximum likelihood estimation
  – we can write down joint distn of all of the $Y$’s, known as the likelihood function

  $$L(\beta_o, \beta_1, \sigma^2 \mid Y_1, \ldots, Y_n) = \prod_{i=1}^{n} N(Y_i \mid \beta_o + \beta_1 x_i, \sigma^2)$$

  – we maximize this to get estimates $\hat{\beta}_o, \hat{\beta}_1$
  – turns out to be equivalent to ....

• Least squares estimation
  – choose $\hat{\beta}_o, \hat{\beta}_1$ to minimize

  $$g(\beta_o, \beta_1) = \sum_{i=1}^{n} (Y_i - (\beta_o + \beta_1 x_i))^2$$

  – least squares has a long history (even without assuming a normal distribution)

  * why squared errors? (convenient math)
  * why vertical distances? ($Y$ is response)

  – result:

  $$\hat{\beta}_o = \bar{Y} - \hat{\beta}_1 \bar{x}$$
  $$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

  – predicted (or fitted) value for a case with $X = x_i$ is

  $$\hat{Y}_i = \hat{\beta}_o + \hat{\beta}_1 x_i$$

  – residual (or error) is $e_i = Y_i - \hat{Y}_i$
Simple linear regression
Estimation - some details

• Least squares estimation:
choose \( \hat{\beta}_o, \hat{\beta}_1 \) to minimize

\[
g(\beta_o, \beta_1) = \sum_{i=1}^{n} (Y_i - (\beta_o + \beta_1 x_i))^2
\]

• Taking derivatives and setting them equal to zero yields normal equations

\[
\begin{align*}
\beta_o n + \beta_1 \sum x_i &= \sum Y_i \\
\beta_o \sum x_i + \beta_1 \sum x_i^2 &= \sum x_i Y_i
\end{align*}
\]

• Solving these equations leads to answers on previous slide
Simple linear regression
Estimation of error variance

• Maximum likelihood estimate of $\sigma^2$ is
  \[ \frac{1}{n} \sum_i (Y_i - \hat{Y}_i)^2 = \frac{1}{n} \sum_i e_i^2 \]

• It turns out that this estimate is generally too small

• A common estimate of $\sigma^2$ is
  \[ s_e^2 = \frac{1}{n - 2} \sum_{i=1}^{n} (Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2 = \frac{1}{n - 2} \sum_{i=1}^{n} e_i^2 \]

  which is used because the $\frac{1}{n-2}$ makes this an unbiased estimate
Simple linear regression
Inference for $\beta_1$

- There are many quantities of interest in a regression analysis
- We may be interested in learning about
  - the slope $\beta_1$
  - the intercept $\beta_o$
  - a particular fitted value $\beta_o + \beta_1 x$
  - a prediction for an individual
- Time is limited so we discuss only drawing statistical conclusions about the slope
Simple linear regression
Inference for the slope, $\beta_1$

- Begin by noting that our estimator of the slope is
  \[ \hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})Y_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \]

- $\hat{\beta}_1$ is a linear combination of normal random variables (the $Y_i$’s) so $\hat{\beta}_1$ is normally distributed
  \[ E(\hat{\beta}_1) = \beta_1 \quad \text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i}(x_i - \bar{x})^2} \]

- $\sigma^2$ is unknown; plug in estimate $s_e^2$

- The estimated standard deviation of $\hat{\beta}_1$ is then
  \[ s_{\beta_1} = \sqrt{s_e^2 / \sum_{i}(x_i - \bar{x})^2} \]

- Then for a large sample size we get an approximate 95% confidence interval for $\beta_1$ is $\hat{\beta}_1 \pm 1.96 s_{\beta_1}$

- More exact confidence interval and test procedures (based on Student’s $t$-distribution) are available but not discussed in this class
Simple linear regression
Model diagnostics - residuals

- We can check whether the linear regression model is a sensible model using the residuals

- Recall $e_i = Y_i - \hat{Y}_i = Y_i - \hat{\beta}_o - \hat{\beta}_1 x_i$

- $e_i$ is an approximation of the stochastic error ($\epsilon_i$) in our model

- Important properties
  - sum of residuals is zero hence a typical value is zero
  - variance of the residuals is approximately equal to one
  - if our model is correct the residuals should look something like a $N(0, 1)$ sample
  - we can look to see if there are patterns in the residuals that argue against our model
Simple linear regression
Diagnosing violations with residual plots

• Plot residuals versus predicted values and look for patterns
  – might detect nonconstant variance
  – might detect nonlinearity
  – might detect outliers

• Histogram or other display of residuals
  – might detect nonnormality

• Show sample pictures in class
Simple linear regression
Remedies for violated assumptions

• What if we find a problem?

• Sometimes the linear regression model will work with a "fix"
  – transform $Y$ (use log $Y$ or $\sqrt{Y}$ as the response)
  – add or modify predictors (perhaps add $X^2$ to approximate a quadratic relationship)

• If no easy "fix", then we can consider more sophisticated models