Hypothesis Testing, Power, Sample Size and Confidence Intervals (Part 1)

B.H. Robbins Scholars Series

June 3, 2010
Outline

Introduction to hypothesis testing
- Scientific and statistical hypotheses
- Classical and Bayesian paradigms
- Type 1 and type 2 errors

One sample test for the mean
- Hypothesis testing
- Power and sample size
- Confidence interval for the mean
- Special case: paired data

One sample methods for a probability
- Hypothesis testing
- Power, confidence intervals, and sample size

Two sample tests for means
- Hypothesis tests
- Power, confidence intervals, and sample size
Introduction

- Goal of hypothesis testing is to rule out chance as an explanation for an observed effect
- Example: Cholesterol lowering medications
  - 25 people treated with a statin and 25 with a placebo
  - Average cholesterol after treatment is 180 with statins and 200 with placebo.
- Do we have sufficient evidence to suggest that statins lower cholesterol?
- Can we be sure that statin use as opposed to a chance occurrence led to lower cholesterol levels?
Hypotheses

▶ **Scientific Hypotheses**

- Often involve estimation of a quantity of interest
- After amputation, to what extent does treatment with clonidine lead to lower rates of phantom limb pain than with standard therapy? (Difference or ratio in rates)
- What is the average increase in alanine aminotransferase (ALT) one month after doubling the dose of medication X? (Difference in means)

▶ **Statistical Hypothesis**

- A statement to be judged. Usually of the form: population parameter X is equal to a specified constant
- Population mean potassium $K$, $\mu = 4.0 \text{ mEq/L}$
- Difference in population means, $\mu_1 - \mu_2 = 0.0 \text{ mEq/L}$
Statistical Hypotheses

- **Null Hypothesis**: $H_0$
  - A straw man; something we hope to disprove
  - It is usually a statement of no effects.
  - It can also be of the form $H_0 : \mu = \text{constant}$, or $H_0 : \text{probability of heads equal 1/2}$.

- **Alternative Hypothesis**: $H_A$
  - What you expect to favor over the null

- If $H_0 : \text{Mean K value} = 3.5 \text{ mEq/L}$
  - One sided alternative hypothesis: $H_A : \text{Mean K} > 3.5 \text{ mEq/L}$
  - Two-sided alternative hypothesis: $H_A : \text{Mean K} \neq 3.5 \text{ mEq/L}$ (values far away from the null)
Classical (Frequentist) Statistics

- Emphasizes hypothesis testing
- Begin by assuming $H_0$ is true
- Examines whether data are consistent with $H_0$
- Proof by contradiction
  - If, under $H_0$, the data are strange or extreme, then doubts are cast on the null.
- Evidence is summarized with a single statistic which captures the tendency of the data.
- The statistic is compared to the parameter value given by $H_0$
Classical (Frequentist) Statistics

▶ **p-value**: Under the assumption that $H_0$ is true, it is the probability of getting a statistic as or more in favor of $H_A$ over $H_0$ than was observed in the data.

▶ Low p-values indicate that if $H_0$ is true, we have observed an improbable event.

▶ Mount evidence against the null, and when sufficient, reject $H_0$.

▶ **NOTE**: Failing to reject $H_0$ does not mean we have gathered evidence in favor of it (i.e., absence of evidence does not imply evidence of absence)
  ▶ There are many reasons for not rejecting $H_0$ (e.g., small samples, inefficient designs, imprecise measurements, etc.)
Classical (Frequentist) Statistics

- Clinical significance is ignored.
- Parametric statistics: assumes the data arise from a certain distribution, often a normal or Gaussian.
- Non-parametric statistics: does not assume a distribution and usually looks at ranks rather than raw values.
Bayesian Statistics

» We can compute the probability that a statement, that is of clinical significance, is true
  » Given the data we observed, does medication X lower the mean cholesterol by more than 10 units?

» May be more natural than the frequentist approach, but it requires a lot more work.

» Supported by decision theory:

» Begin with a (prior) belief → learn from your data → Form a new (posterior) belief that combines the prior belief and the new data

» We can then formally integrate information accrued from other studies as well as from skeptics.

» Becoming more popular.
Errors in Hypothesis Testing

- Type 1 error: Reject $H_0$ when it is true
  - Significance level ($\alpha$) or Type 1 error rate: is the probability of making this type of error
  - This value is usually set to 0.05 for random reasons

- Type 2 error: Failing to reject $H_0$ when it is false
  - The value $\beta$ is the probability of a type 2 error or type 2 error rate.

- Power: $1 - \beta$: probability of correctly rejecting $H_0$ when it is false

<table>
<thead>
<tr>
<th>Decision</th>
<th>State of $H_0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Do not reject $H_0$</td>
<td>$H_0$ is true</td>
</tr>
<tr>
<td>Reject $H_0$</td>
<td>Type 1 error ($\alpha$)</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Notes Regarding Hypothesis Testing

- **Two schools of thought**
  - Neyman-Pearson: Fix Type 1 error rate (say $\alpha = 0.05$) and then make the binary decision, reject/do not reject
  - Fisher: Compute the p-value and quote the report in the publication.
  - We favor Fisher, but Neyman-Pearson is used all of the time.

- **Fisher approach:** discussion of p-values does not require discussion of type 1 and type 2 errors
  - Assume the sample was chosen randomly from a population whose parameter value is captured by $H_0$. The p-value is a measure of evidence against it.

- **Neyman-Pearson approach:** having to make a binary call (reject vs do not reject) regarding significance is arbitrary
  - There is nothing magical about 0.05
  - Statistical significance has nothing to do with clinical significance
One sample test for the mean

- Assumes the sample is drawn from a population where values are normally distributed (normality is actually not necessary)
- One sample tests for mean $\mu = \mu_0$ (constant) don’t happen very often except when data are paired (to be discussed later)
- The t-test is based on the t-statistic

$$t = \frac{\text{estimated value} - \text{hypothesized value}}{\text{standard deviation of numerator}}$$

- Standard deviation of a summary statistic is called the standard error which is the square root of the variance of the statistic
One sample test for the mean

- Sample average: $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$
  - The estimate of the population mean based on the observed sample
- Sample variance: $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$
- Sample standard deviation: $s = \sqrt{s^2}$
- $H_0 : \mu = \mu_0$ vs. $H_A : \mu \neq \mu_0$
- One sample t-statistic
  \[ t = \frac{\bar{x} - \mu_0}{SE} \]
- Standard error of the mean, $SE = \frac{s}{\sqrt{n}}$
Hypothesis Testing, Power, Sample Size and Confidence Intervals (Part 1)

One sample test for the mean

Hypothesis testing

One sample t-test for the mean

- When data come from a normal distribution and $H_0$ holds, the $t$ ratio follows the $t-$ distribution. What does that mean?
- Draw a sample from the population, conduct the study and calculate the t-statistic.
- Do it again, and calculate the t-statistic again.
- Do it again and again.
- Now look at the distribution of all of those t-statistics.
- This tells us the relative probabilities of all t-statistics if $H_0$ is true.
Example: one sample t-test for the mean

- The distribution of potassium concentrations in the target population are normally distributed with mean 4.3 and variance 0.1: $N(4.3, .1)$.

- $H_0 : \mu = 4.3$ vs. $H_A : \mu \neq 4.3$. Note that $H_0$ is true!

- Each time the study is done,
  - Sample 100 participants
  - Calculate:

$$t = \frac{\bar{x} - 4.3}{SE}$$

- Conduct the study 25 times, 250 times, 1000 times, 5000 times
Hypothesis Testing, Power, Sample Size and Confidence Intervals (Part 1)

- One sample test for the mean
- Hypothesis testing
One sample t-test for the mean

- With very small samples (n), the t statistic can be unstable because the sample standard deviation (s) is not a precise estimate of the population standard deviation (σ).
- So, the t-statistic has heavy tails for small n.
- As n increases, the t-distribution converges to the normal distribution with mean equal to 0 and with standard deviation equal to one.
- The parameter defining the particular t-distribution we use (function of n) is called the degrees of freedom or d.f.
- d.f. = n - number of means being estimated
- For the one-sample problem, d.f. = n - 1
- Symbol is $t_{n-1}$
Hypothesis Testing, Power, Sample Size and Confidence Intervals (Part 1)

One sample test for the mean

Hypothesis testing

Density for the t-distribution

- t (d.f.=5)
- t (d.f.=10)
- t (d.f.=100)
- N (0,1)
One sample t-test for the mean

- One sided test: $H_0 : \mu = \mu_0$ versus $H_A : \mu > \mu_0$
- One tailed p-value:
  - Probability of getting a value from the $t_{n-1}$ distribution that is at least as much in favor of $H_A$ over $H_0$ than what we had observed.
- Two-sided test: $H_0 : \mu = \mu_0$ versus $H_A : \mu \neq \mu_0$
- Two-tailed p-value:
  - Probability of getting a value from the $t_{n-1}$ distribution that is at least as big in absolute value as the one we observed.
One sample t-test for the mean

- Computer programs can compute the p-value for a given n and t-statistic
- Critical value
  - The value in the t (or any other) distribution that, if exceeded, yields a ‘statistically significant’ result for type 1 error rate equal to $\alpha$
- Critical region
  - The set of all values that are considered statistically significantly different from $H_0$. 
One sample test for the mean

Hypothesis testing

T-distribution (d.f.=10) and one-sided critical region (0.05)

T-distribution (d.f.=100) and one-sided critical region (0.05)

T-distribution (d.f.=10) and two-sided critical regions (0.05)

T-distribution (d.f.=100) and two-sided critical regions (0.05)
Power and Sample Size for a one sample test of means

- Power increases when
  - Type 1 error rate ($\alpha$) increases: type 1 ($\alpha$) versus type 2 ($\beta$) tradeoff
  - True $\mu$ is very far from $\mu_0$
  - Variance or standard deviation ($\sigma$) decreases (decrease noise)
  - Sample size increases

- T-statistic

$$t = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

- Power for a 2-tailed test is a function of the true mean $\mu$, the hypothesized mean $\mu_0$, and the standard deviation $\sigma$ only through $|\mu - \mu_0|/\sigma$
Sample size to achieve $\alpha = 0.05$, power=0.90 is approximately

$$n = 10.51 \left( \frac{\sigma}{\mu - \mu_0} \right)^2$$

Power calculators can be found at statpages.org/#Power

PS is a very good power calculator (Dupont and Plummer): http://biostat.mc.vanderbilt.edu/PowerSampleSize
Example: Power and Sample Size

▶ The mean forced expiratory volume in 1 second in a population of asthmatics is 2.5 L/sec, and the standard deviation is assumed to be 1

▶ How many subjects are needed to reject $H_0 : \mu = 2.5$ in favor of $H_0 : \mu \neq 2.5$ if the new drug is expected to increase the FEV to 3 L/sec with $\alpha = 0.05$ and $\beta = 0.1$

▶ $\mu_0 = 2.5, \mu = 3.0, \sigma = 1$

\[
n = 10.51 \left( \frac{1}{3.0 - 2.5} \right)^2 = 42.04
\]

▶ We need 43 subjects to have 90 percent power to detect a 0.5 difference from 2.5.
Confidence Intervals

- Two-sided, $100(1 - \alpha)\%$ CI for the mean $\mu$ is given by
  \[(x - t_{n-1,1-\alpha/2} \cdot SE, x + t_{n-1,1-\alpha/2} \cdot SE)\]
- $t_{n-1,1-\alpha/2}$ is the critical value from the t-distribution with d.f.$=n-1$
- For large $n$, $t_{n-1,1-\alpha/2}$ is equal to 1.96 for $\alpha = 0.05$
- $1 - \alpha$ is called the confidence level or confidence coefficient
Confidence Intervals

- **100(1 − α)% confidence interval (CI)**
  - If we were able to repeat a study a large number of times, then $100 \cdot (1 - \alpha)$ percent of CIs would contain the true value.

- **Two-sided 100(1 − α)% CI**
  - Includes the null hypothesis $\mu_0$ if and only if a hypothesis test $H_0 : \mu = \mu_0$ is not rejected for a 2-sided $\alpha$ significance level test.
  - If a 95% CI does not contain $\mu_0$, we can reject $H_0 : \mu = \mu_0$ at the $\alpha = 0.05$ significance level.

<table>
<thead>
<tr>
<th>n</th>
<th>$\bar{x}$</th>
<th>$\sigma$</th>
<th>p-value</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>27.31</td>
<td>54.23</td>
<td>0.036</td>
<td>(1.930, 52.690)</td>
</tr>
<tr>
<td>20</td>
<td>27.31</td>
<td>59.23</td>
<td>0.053</td>
<td>(-0.410, 55.030)</td>
</tr>
<tr>
<td>20</td>
<td>25.31</td>
<td>54.23</td>
<td>0.051</td>
<td>(-0.070, 50.690)</td>
</tr>
<tr>
<td>17</td>
<td>27.31</td>
<td>54.23</td>
<td>0.054</td>
<td>(-0.572, 55.192)</td>
</tr>
</tbody>
</table>

- CIs provide more information than p-values
Special case: Paired data and one-sample tests

- Assume we want to study whether furosemide (or lasix) has an impact on potassium concentrations among hospitalized patients.
- That is, we would like to test $H_0 : \mu_{on-furo} - \mu_{off-furo} = 0$ versus $H_A : \mu_{on-furo} - \mu_{off-furo} \neq 0$
- In theory, we could sample $n_1$ participants not on furosemide and compare them to $n_2$ participants on furosemide.
- However, a very robust and efficient design to test this hypothesis is with a paired sample approach.
- On $n$ patients, measure K concentrations just prior to and 12 hours following furosemide administration.
The effect measure to test $H_0$ versus $H_A$, is the mean, within person difference between pre and post- administration K concentrations.

$W_i = Y_{on-furo,i} - Y_{off-furo,i}$

Note that $W = \overline{Y}_{on-furo} - \overline{Y}_{off-furo}$

The average of the differences is equal to the difference between the averages.

$H_0: \mu_w = 0$ versus $H_A: \mu_w \neq 0$ is equivalent to the above $H_0$ and $H_A$.

$\overline{W} = -0.075$ mEq/L and $s = 0.08$

$$t_{99} = \frac{-0.075 - 0}{0.08/\sqrt{100}} = 9.375$$

The p-value is less than 0.0001 → a highly (!!!) statistically significant reduction.
One Sample Methods for a Probability

- **Y** is binary (0/1): Its distribution is bernoulli($p$) ($p$ is the probability that $Y = 1$).
- $p$ is also the mean of $Y$ and $p(1 - p)$ is the variance.
- We want to test $H_0 : p = p_0$ versus $H_A : p \neq p_0$
- Estimate the population probability $p$ with the sample proportion or sample average $\hat{p}$

\[
\hat{p} = \frac{1}{n} \sum_{i=1}^{n} Y_i
\]
One Sample Methods for a Probability

- A z-test is an approximate test that assumes the test statistic has a normal distribution i.e., it is a t-statistic with the d.f. very large

\[ z = \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}} \]

- The z-statistic has the same form as the t-statistic

\[ z = \frac{\text{estimated value} - \text{hypothesized value}}{\text{standard deviation of numerator}} \]

where \( \sqrt{p_0(1 - p_0)/n} \) is the standard deviation of the numerator which is the standard error assuming the \( H_0 \) is true.

- (see t-statistic distributions)
One Sample test for a probability: Is our coin fair?

- $Y \sim \text{bernoulli}(p)$: $H_0 : p = 0.5$ versus $H_A : p \neq 0.5$
- Flip the coin 50 times. Heads ($Y=1$) shows up 30 times ($\hat{p} = 0.6$).

$$z = \frac{0.6 - 0.5}{\sqrt{(0.5)(0.5)/50}} = 1.414$$

- The p-value associated with $Z$ is $2 \times$ the area under the normal curve to the right of $z=1.414$ (e.g. the area to the right of 1.414 plus the area to the left of -1.414)
- The critical value for a 2-sided $\alpha = 0.05$ significance level test is 1.96
- The p-value associated with this test is approximately 0.16
- Note that if $p$ is very small or very large or if $n$ is small, use exact methods (e.g. Fishers exact test or permutation test)
Z-test for a proportion: Z-statistic=1.414
Power and confidence intervals

- Power increases when
  - $n$ increases
  - $p$ departs from $p_0$
  - $p_0$ departs from 0.5

\[ z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} \]

- Confidence interval
  - 95% CI: $(\hat{p} - 1.96 \cdot \sqrt{\hat{p}(1-\hat{p})/n}, \hat{p} - 1.96 \cdot \sqrt{\hat{p}(1-\hat{p})/n})$

- For the coin flipping example: $\hat{p} = 0.6$ and the 95% CI is given by

\[ 0.6 \pm 1.96 \cdot \sqrt{0.6 \times 0.4/50} = (0.464, 0.736) \]

which is consistent with the 0.16 p-value that we had observed for $H_0 : p = 0.5$. 

Two sample test for means

- Two groups of patients (not paired)
- These are much more common than 1 sample tests
- We assume data come from a normal distribution (although this is not completely necessary)
- For now, assume the two groups have equal variability in response distribution
- Test whether population means are equal
- Example: All patient in population 1 are treated with clonidine after limb amputation and all patients in population 2 are treated with standard therapy.
- Scientific question:
  - What is the difference in the mean pain scale scores at 6 months following the amputation?
Two sample test for means

- $H_0 : \mu_1 = \mu_2$ which can be generalized to $H_0 : \mu_1 - \mu_2 = 0$
or $H_0 : \mu_1 - \mu_2 = \delta$

- The quantity of interest (QOI) is $\mu_1 - \mu_2$

- If we want to test $H_0 : \mu_1 - \mu_2 = 0$ and if we assume the two populations have equal variances, then the $t$-statistic is given by:

\[
t = \frac{\text{point estimate of the QOI} - 0}{\text{standard error of the numerator}}
\]

- The estimate of the QOI: $\bar{x}_1 - \bar{x}_2$
Two sample test for means

- For two independent samples variance of the sum or of differences in means is equal to the sum of the variances
- The variance of the QOI is then given by $\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}$
- We need to estimate a single $\sigma^2$ from the two samples
- We use a weighted average of the two sample variances

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

- The true standard error of the difference in sample means:

$$\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

- Estimate with $s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$
Two sample test for means

- The t-statistic is given by,

\[ t = \frac{\bar{x}_1 - \bar{x}_2}{s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \]

- Under \( H_0 \), \( t \) has a t-distribution with \( n_1 + n_2 - 2 \) degrees of freedom.

- The -2 comes from the fact that we had to estimate the center of 2 distributions.
Example: two sample test for means

- \( n_1 = 8, \; n_2 = 21, \; s_1 = 15.34, \; s_2 = 18.23, \; \bar{x}_1 = 132.86, \; \bar{x}_2 = 127.44 \)

\[
s^2 = \frac{7(15.34)^2 + 20(18.23)^2}{7 + 20} = 307.18
\]

\[
s = \sqrt{307.18} = 17.527
\]

\[
se = 17.527 \sqrt{\frac{1}{8} + \frac{1}{21}} = 7.282
\]

\[
t = \frac{5.42}{7.282} = 0.74
\]

on 27 d.f.
Example: two sample test for means

- The two-sided p-value is 0.466
  - You may verify with the surfstat.org t-distribution calculator
- The chance of getting a difference in means as large or larger than 5.42 if the two populations have the same mean in 0.466.
- No evidence to suggest that the population means are different.
Power and sample size: two sample test for means

- Power increases when
  - $\Delta = |\mu_1 - \mu_2|$ increases
  - $n_1$ or $n_2$ increases
  - $n_1$ and $n_2$ are close
  - $\sigma$ decreases
  - $\alpha$ increases

- Power depends on $n_1$, $n_2$, $\mu_1$, $\mu_2$, and $\sigma$ approximately through

$$\frac{\Delta}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

- When using software to calculate power you can put in 0 for $\mu_1$ and $\Delta$ for $\mu_2$ since all that matters is their difference
- $\sigma$ is often estimated from pilot data
Power and sample size: two sample test for means

Example

- From available data, ascertain a best guess of $\sigma$: assume it is 16.847.
- Assume $\Delta=5$, $n_1 = 100$, $n_2 = 100$, $\alpha = 0.05$
- The surfstat software computes a power of 0.555

The required sample size decreases with

- $k = \frac{n_2}{n_1} \to 1$
- $\Delta$ large
- $\sigma$ small
- $\alpha$ large
- Lower power requirements
Power and sample size: two sample test for means

- An approximate formula for required sample sizes to achieve power=0.9 with $\alpha = 0.05$ is

$$n_1 = \frac{10.51\sigma^2(1 + \frac{1}{k})}{\Delta^2}$$

$$n_2 = \frac{10.51\sigma^2(1 + k)}{\Delta^2}$$

<table>
<thead>
<tr>
<th>$\sigma$</th>
<th>$\Delta$</th>
<th>$K$</th>
<th>$n_1$</th>
<th>$n_2$</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>16.847</td>
<td>5</td>
<td>1.0</td>
<td>239</td>
<td>239</td>
<td>478</td>
</tr>
<tr>
<td>16.847</td>
<td>5</td>
<td>1.5</td>
<td>199</td>
<td>299</td>
<td>498</td>
</tr>
<tr>
<td>16.847</td>
<td>5</td>
<td>2.0</td>
<td>177</td>
<td>358</td>
<td>537</td>
</tr>
<tr>
<td>16.847</td>
<td>5</td>
<td>3.0</td>
<td>160</td>
<td>478</td>
<td>638</td>
</tr>
</tbody>
</table>

- Usually, websites are recommended for these calculations.
Confidence interval: two sample test for means

- Confidence interval

\[
\left[ (\bar{x}_1 - \bar{x}_2) - t_{n_1+n_2-2,1-\alpha/2} \times s \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, \right.
\]

\[
\left. (\bar{x}_1 - \bar{x}_2) + t_{n_1+n_2-2,1-\alpha/2} \times s \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right]
\]

<table>
<thead>
<tr>
<th>( \Delta )</th>
<th>( s )</th>
<th>( n_1 )</th>
<th>( n_2 )</th>
<th>LCI</th>
<th>UCI</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>16.847</td>
<td>100</td>
<td>100</td>
<td>3.01</td>
<td>6.99</td>
</tr>
<tr>
<td>5</td>
<td>16.847</td>
<td>75</td>
<td>125</td>
<td>2.95</td>
<td>7.05</td>
</tr>
<tr>
<td>5</td>
<td>16.847</td>
<td>50</td>
<td>150</td>
<td>2.70</td>
<td>7.30</td>
</tr>
</tbody>
</table>
Summary

- Hypothesis testing, power, sample size, and confidence intervals
  - One sample test for the mean
  - One sample test for a probability
  - Two sample test for the mean