The Role of Chance in Statistical Testing

Shahidul Islam, MPH, PStat®
Department of Biostatistics
Disclaimer

- **This is not** an introduction to probability and statistics

- **This is** to give you working knowledge of why and how we use statistics/probability in medicine/epidemiology
Outline

• A quick review of probability
• Background of experiments
• Sampling variability and Random error
• P-values
• Confidence intervals
• Power
Probability

• Simple Experiment#1: Toss a coin. What is the probability of obtaining a head?
  – Possible outcomes (sample space) = \{H, T\}
  – So, number of events in this case \(n(E)=1\) and number of all possible outcomes \(n(S)=2\)
  – The probability \(P=n(E)/n(S)\) i.e. \(\frac{1}{2}\)

• Simple experiment#2: Toss two coins, what is the probability of obtaining two heads?
  – Sample space \(S=\{HT, HH, TH, TT\}\)
  – Event \(E=HH\)
  – Probability of two heads=\(n(E)/n(S)=1/4\)
• Experiment#3: A die is rolled, find the probability that an even number is obtained.
  – Sample Space $S=\{1, 2, 3, 4, 5, 6\}$
  – Event $E=\{2, 4, 6\}$
  – $P(E) = \frac{n(E)}{n(S)} = \frac{3}{6} = \frac{1}{2}$ i.e. 0.5

• Experiment#4: Two dice are rolled, find the probability that the sum is equal to 4
  – $S = \{(1,1),(1,2),(1,3),(1,4),(1,5),(1,6), (2,1),(2,2),(2,3),(2,4),(2,5),(2,6), (3,1),(3,2),(3,3),(3,4),(3,5),(3,6), (4,1),(4,2),(4,3),(4,4),(4,5),(4,6), (5,1),(5,2),(5,3),(5,4),(5,5),(5,6), (6,1),(6,2),(6,3),(6,4),(6,5),(6,6)\}$
  – $E=\{(1,3),(2,2),(3,1)\}$
  – $P(E) = \frac{3}{36} = \frac{1}{12}$
Basic rules of probability

• P(E) must be between 0 and 1.
  – If event A can never occur, P(A) = 0.
  – If event A always occurs when the experiment is performed, P(A) = 1.

• The sum of the probabilities for all simple events in Sample space equals 1.
The first ever clinical trial?

- James Lind was a Scottish physician who was a pioneer of naval hygiene in the Royal Navy.

- Scurvy (a disease resulting from a lack of vitamin C) was killing many sailors when he joined the Navy.

- Lind thought that scurvy was due to deterioration of the body which could be helped by acids. So, in 1747, he conducted a trial.

- He divided 12 sailors into six groups of two (A six-armed trial)
  - They all received the same diet
  - 5 groups of two received elixir vitriol, sea water, cider, vinegar or spice mix
  - One group of two received lemons and oranges
Output and conclusion of Lind’s trial

- None of ten sailors receiving elixir vitriol, sea water, cider, vinegar or spice mix showed any improvement.

- Both sailors who received lemons and oranges were cured.

- Still, Lind concluded that the study was too small and needed to be repeated by other scientists.

- Was he right?
Possible explanation

• **True success**: Treatment success Lind observed was true

• **Bias**: A systematic error in the conduct of the study which offsets the results in some non-random way

• **Confounding**: Mixing of the effect of the exposure and the effect of some other variable

• **Chance**: Random distortion of the results due to sampling variability
Possible outcome in a study

A study is Conducted

Association Observed

True Association

Random Error

Systematic Error

Insufficient Sample Size

Systematic Error

Truly No association

No Association observed

Bias

Confounding

Bias

Confounding
## Random errors

<table>
<thead>
<tr>
<th></th>
<th>H0 is True</th>
<th>H0 is False</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reject H0</td>
<td>Type-I Error</td>
<td>Correct Decision</td>
</tr>
<tr>
<td>Do not reject H0</td>
<td>Correct Decision</td>
<td>Type-II Error</td>
</tr>
</tbody>
</table>

Usually, Type-I error is set at 0.05 (Alpha or significance level and Type-II error is set at 0.20 (Beta)
Sampling variability

- This glass jar contains 60 balls (30 red and 30 blue)
- Let say, we would like to find out the proportion of red balls without having to count them all.
Sampling variability

Proportion of red balls:

- Sample 1: 50% = Correct
- Sample 2: 0% = Wrong
- Sample 3: 75% = Wrong
Sampling variability

• Can we use statistics to guide our evaluation of the random sample?
  – Yes, statistical inference

• Sample insights
  – The larger the sample, the smaller the sampling variability
  – Larger samples decrease the probability of drawing unrepresentative samples
• Inference from statistical samples are often based on the following:
  – An hypothesis is specified
  – A sample is drawn
  – An appropriate statistical test is performed to quantify the statistical uncertainty associated with the sample to test the hypothesis
The following data was collected in a clinical trial evaluating a new compound designed to improve wound healing in trauma patients. The new compound is compared against a placebo. After treatment for 5 days with the new compound or placebo, the extent of wound healing is measured and the data are shown below. The number in each cell is a count.

<table>
<thead>
<tr>
<th>Percent Wound Healing</th>
<th>0-25%</th>
<th>26-50%</th>
<th>51-75%</th>
<th>76-100%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>New Compound (n=125)</td>
<td>15</td>
<td>37</td>
<td>32</td>
<td>41</td>
</tr>
<tr>
<td>Placebo (n=125)</td>
<td>36</td>
<td>45</td>
<td>34</td>
<td>10</td>
</tr>
</tbody>
</table>

Is there a difference in the extent of wound healing by treatment? (Hint: Are treatment and the percent wound healing independent?) Run the appropriate test at a 5% level of significance.
Statistical Inference

• This scenario is a good fit for the Chi-Square test because both the ‘treatment’ and ‘percent wound healing’ are categorical variables. We are interested in evaluating whether these two variables are associated.

• So, let’s follow these steps to apply chi-square test to solve the above problem.
Step 1. Set up hypotheses and determine level of significance.

Null hypothesis \((H_0)\): Treatment and percent wound healing are independent (not associated)

Alternative hypothesis \((H_1)\): \(H_0\) is false.

Significance level \(\alpha=0.05\)  

Type-I error: This is the accepted probability of rejecting null hypothesis when it is true
Step 2. Select the appropriate test statistic formula for Chi-Square test.

\[ \chi^2 = \sum \frac{(O - E)^2}{E} \]

Where,
O=Observed count
E=Expected count

The condition for appropriate use of the above test statistic is that each expected frequency is at least 5.

In Step 4 we will compute the expected frequencies and we will ensure that the condition is met.
Statistical Inference

**Step 3.** Set up decision rule.
Degrees of freedom (D.F.) = (number of rows – 1) (number of columns – 1)

D.F.=(2-1)(4-1)=3 and the decision rule is Reject H₀ if $\chi^2 \geq 7.81$.

**Step 4.** Compute the test statistic.

We now compute the expected frequencies using the formula,
- Expected Frequency = (Row Total * Column Total)/N.
### Statistical Inference

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Percent Wound Healing</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0-25%</td>
<td>26-50%</td>
</tr>
<tr>
<td>New Compound</td>
<td>15</td>
<td>37</td>
</tr>
<tr>
<td></td>
<td>(25.5)</td>
<td>(41)</td>
</tr>
<tr>
<td>Placebo</td>
<td>36</td>
<td>45</td>
</tr>
<tr>
<td></td>
<td>(25.5)</td>
<td>(41)</td>
</tr>
<tr>
<td>Total</td>
<td>51</td>
<td>82</td>
</tr>
</tbody>
</table>

\[
\chi^2 = \frac{(15 - 25.5)^2}{25.5} + \frac{(37 - 41)^2}{41} + \frac{(32 - 33)^2}{33} + \frac{(41 - 25.5)^2}{25.5} + \frac{(36 - 25.5)^2}{25.5} + \frac{(45 - 41)^2}{41} + \frac{(34 - 33)^2}{33} + \frac{(10 - 25.5)^2}{25.5}
\]

\[\chi^2 = 4.32 + 0.39 + 0.03 + 9.42 + 4.32 + 0.39 + 0.03 + 9.42 = 28.32\]
Step 5. Conclusion.

Reject the null hypothesis \((H_0)\) because test statistic \(28.32 > 7.81\).

The \textbf{P value} is \(\chi^2 (3) = 28.32, P<0.001\).

We have enough evidence at 5% level of significance \((\alpha=0.05)\) to reject null hypothesis \((H_0)\) and conclude that treatment and wound healing are associated.

We also conclude that chance is an unlikely explanation for the observed association.
$P$ Value

• The infamous $P$ value is the probability that the observed results could have been due to chance, given that the null hypothesis is true.

• As we have seen, on the basis of the $P$ value, we either reject or accept the null hypothesis.
Notes $P$ values

• Please note that defect in study design makes an interpretation of a $P$ value useless

• Significant $P$ values do not exclude the possibility of bias and confounding issues
  – Bias is systematic error during designing and conducting of a study that results in invalid estimate of the measures of association.
    • Most common biases: Selection bias, recall bias, information bias
  – Study found that significantly more coffee drinkers die younger compared to non-coffee drinkers, can you think of a confounder?

• We need to consider bias and confounding factors as possible explanations of any finding before interpreting a p-value
• If a test of the null hypothesis gave $P=0.01$, the null hypothesis has only a 1% chance of being true; if instead it gave $P=0.40$, the null hypothesis has a 40% chance of being true.

– No, $P=0.01$ would indicate that the data are not very close to the null hypothesis and $P=0.40$ would indicate that the data are much closer to the null hypothesis i.e. chance is a likely explanation for the observed effect.
Common misinterpretations of *P* values

- **Statistical significance (**$P<0.05$**)** indicates a scientifically or substantively important relation has been detected.
  - No! Especially when a study is large, very minor effects or small assumption violations can lead to statistically significant *p*-value.

- **Lack of statistical significance (**$P>0.05$**)** indicates that the effect size is small.
  - No! Especially when a study is small, even large effects may be "drowned in noise" and thus fail to be detected as statistically significant.
Example: Difference between Clinical and Statistical Significance

<table>
<thead>
<tr>
<th>Drug</th>
<th>N</th>
<th>BP mean decrease</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agent #1</td>
<td>500</td>
<td>30 mm Hg</td>
<td>15</td>
</tr>
<tr>
<td>Agent #2</td>
<td>500</td>
<td>32 mm Hg</td>
<td>16</td>
</tr>
</tbody>
</table>

Using unpaired t-test, \( p = 0.041 \)

What’s Wrong?

The statistically significant result is due to large samples. Difference of 2 mm is not clinically important.
Difference between Clinical and Statistical Significance

<table>
<thead>
<tr>
<th>Drug</th>
<th>N</th>
<th>BP mean decrease</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agent #1</td>
<td>10</td>
<td>30 mm Hg</td>
<td>15</td>
</tr>
<tr>
<td>Agent #2</td>
<td>10</td>
<td>17 mm Hg</td>
<td>16</td>
</tr>
</tbody>
</table>

Using unpaired t-test, \( p = 0.077 \)

**What’s Wrong?**

Although statistical significance was not quite reached, the difference of 13 mm Hg is clinically relevant.
Significance level

• We commonly use the significance threshold level of 5% (0.05)

• At or below this arbitrary level,
  – We reject the null hypothesis
  – Consider our findings “statistically significant”
  – Conclude that “chance is an unlikely explanation for the observed difference”
Significance level

• At the most common significance level of alpha=0.05, we make sure that type-I error rate is only 5%
  
  – For every 100 hypothesis tests performed, 5 will be “false positive”

  – Is this acceptable?

  – What about other significance level such as alpha=0.01?
Confidence Intervals

• Confidence intervals represent an alternative way of evaluating hypothesis tests

• A confidence interval describes the sampling variability of the point estimate

• As with $P$ values, confidence intervals are constructed and evaluated with a type-I error rate in mind (usually 5% i.e. alpha=0.05)
Confidence Intervals

- If you conducted your study 100 times you would produce 100 different confidence intervals. We would expect that 95 out of those 100 confidence intervals will contain the true population mean.

- Example CI: In the Framingham Offspring Study (n=3534), the mean systolic blood pressure (SBP) was 127.3 with a standard deviation of 19.0. Generate a 95% confidence interval for the true mean SBP.

\[
\bar{X} \pm Z \frac{s}{\sqrt{n}} = 127.3 \pm 1.96 \frac{19.0}{\sqrt{3534}}
\]

\[
127.3 \pm 0.63 \quad (126.7, 127.9)
\]
Confidence Intervals

• Width of the interval inform us about the precision of the point estimate
  – Confidence intervals more versatile and useful than p-values

• Confidence interval can also be used for hypothesis testing by examining whether the interval excludes the null hypothesis
  – Such as Relative Risk=1, Odds Ratio=1, Risk difference=0
Confidence Interval

• Examples

<table>
<thead>
<tr>
<th></th>
<th>Hypertensive</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Yes</td>
</tr>
<tr>
<td>Experimental</td>
<td>14</td>
</tr>
<tr>
<td>Placebo</td>
<td>22</td>
</tr>
<tr>
<td>Total</td>
<td>36</td>
</tr>
</tbody>
</table>

Relative Risk (RR) = \( \frac{14/100}{22/100} \)
= \( \frac{0.14}{0.22} = 0.64 \)

RR=0.64, i.e. experimental drug reduced the risk of hypertension by 36%. Is it statistically significant?

95% Confidence Interval for RR=0.64 is \([0.35, 1.17]\)
Common misinterpretations of confidence intervals

• There is a 95% chance that the true population mean falls within the confidence interval presented by a study
  – NO, it really means that if you conducted your study 100 times you would produce 100 different confidence intervals. We would expect that 95 out of those 100 confidence intervals will contain the true population mean.

• If two confidence intervals overlap, the difference between two estimates or studies is not significant
  – NO, it is always true that if the confidence intervals do not overlap, then the statistics will be statistically significantly different. However, the converse is not true. The CI can overlap by about 25% and it still can be significant
Type-II error and Power

• Type-II error i.e. we conclude that there is no difference when there really is a difference

• Type-II errors usually occurs when the sample size is too small i.e. when we have insufficient power
  – Power=1-Type II error = 1-Beta
Alpha(\(\alpha\)), Beta(\(\beta\)) and Power

Distribution of diastolic BP under H0: mean=90 and under H1: mean=94
Statistical Power

• Power computations can help us design a study to determine
  – How many participants we need to be able to detect the desired effect size
  – We can also use the power calculations to determine the effect sizes we can detect given a certain sample size

• Beta (β) and Power are related to the sample size, level of significance (α) and the effect size (difference in parameter of interest under H₀ versus H₁)
  – Power is higher with larger alpha (α)
  – Power is higher with larger effect size
  – Power is higher with larger sample size
Find sample size to test H0: \( \mu = \mu_0 \)

- Planning study to test the following hypothesis for mean expenditures
  
  \( H_0: \mu = $3302 \) vs.
  
  \( H_1: \mu \neq $3302 \) at \( \alpha = 0.05 \)

- Determine sample size to ensure 80% power to detect a difference of $150 in mean expenditures on health care and prescription drugs (assume standard deviation is $890).
Find sample size to test H0: $\mu = \mu_0$

\[ ES = \frac{|\mu_1 - \mu_0|}{\sigma} = \frac{150}{890} = 0.17 \]

\[ n = \left( \frac{Z_{1-\alpha/2} + Z_{1-\beta}}{ES} \right)^2 = \left( \frac{1.96 + 0.84}{0.17} \right)^2 = 271.3 \]

Need a sample size of 272 to detect $150 difference in health care expenditure with 80% power
Find sample size to test $H_0: p = p_0$

- Planning study to test
  
  $H_0: p = 0.26$ vs.
  
  $H_1: p \neq 0.26$ at $\alpha = 0.05$

- Determine sample size to ensure 90% power to detect a difference of 5% in the proportion of patients with elevated LDL cholesterol.
Need sample size of 869 to detect a difference of 5% in the proportion of patients with elevated LDL cholesterol with 90% power.
To sum up...

• P-values and confidence intervals are the key tools in statistical hypothesis testing
  
  – A p-value provides the probability that the observed result could have arisen due to chance, given the null hypothesis is true

  – A confidence interval provides the range within which the true point estimate lies with a certain probability (usually 95%)
To sum up...

• In addition to the effect of bias and confounding, chance must always be taken into consideration as a cause of an observed association

• Conventionally, we use 5% significance level, which translate to 5% false positive rate in any given hypothesis testing

• While useful and powerful, p-values say
  – Nothing about the possibility of bias
  – Nothing whatsoever about causality
Department of Biostatistics
Biostatistics@nyulangone.org