Hypothesis Testing & Central Limit Theorem in R

Paul Stey

January 9, 2018

Table of Contents

Four Books

- "*Statistics as Principled Argument*", Abelson, R.P., 1995
- "*Discovering Statistics using R*", Field, A. *et al.*, 2012
- "*Statistical Rethinking*", McElreath, R., 2015
- "*All of Statistics*", Wasserman, L., 2004

Bayesians, Frequentists, and Likelihoodists

There are a few approaches to statistical inference:

1. Bayesian
2. Likelihoodist
3. Frequentist

We will be concerned primarily with the frequentist approach; but we recommend you take time to explore Bayesian methods (to update your beliefs ☺).

What is hypothesis testing?

Hypothesis testing is the process of using data to make decisions under uncertainty.

What is hypothesis testing? (*cont.*)

The frequentist approach is

1. Typically choosing between 2 competing hypotheses.
   - Null hypothesis (usually written H_0)
   - Alternative hypothesis (usually written H_1 or sometimes H_A)

What is hypothesis testing? *cont.*

For example, we might be interested in whether some new medication, M, reduces cholesterol. Here the competing hypotheses are:

H_0: $\mu_1 = \mu_2$ "M *does not reduce cholesterol*" (null hypothesis)
H_1: $\mu_1 < \mu_2$ "M *reduces cholesterol*" (alternative hypothesis)

where μ_1 is mean cholesterol for those receiving M in the population, and μ_2 is mean cholesterol for those *not* receiving M in the population.

Notes on hypothesis testing

Some important things to note:

1. Previous example is one-sided test; two-sided tests generally look like:
   - H_0: $\mu_1 = \mu_2$
   - H_1: $\mu_1 \neq \mu_2$
2. Two-sided tests tend to be more common
3. You should clearly articulate hypotheses *priori to conducting statistical tests*

Notes on Hypothesis Testing *cont.*

General process of hypothesis testing:

1. Specify the null and alternative hypotheses, H_0 and H_1
2. Determine the test to be used, which gives us:
   - Our test statistic
   - Corresponding probability distribution
3. Set a level of significance (e.g., $\alpha = 0.05$)
4. Use our data to compute our test statistic (and perhaps its standard error)
5. Use test statistic and its accompanying distribution to obtain *p*-value

p-values

What is a p-value?

p-values *cont.*

A *p*-value is a probability. In particular, it is the probability of finding data *as extreme or more extreme* than what he have observed, given that the null hypothesis is true.

p-values *cont.*

In other words, a *p*-value can be used to answer this question:

"*If the null hypothesis is true, are my data unusual?*"

When a *p*-value is small, our answer is "*yes*". And when the answer is "*yes*", we are generally inclined to take this as evidence against the null hypothesis.

p-values *cont.*

A *p*-value is **NOT**:

1. The probability the null hypothesis is true
2. The probability that the data were produced by chance alone
3. A measure of effect size
   - Be wary of papers discussing "highly" or "extremely" significant results based *p*-values
   - Also beware of studies using *p*-values as inputs to subsequent computations or tests

p-values *cont.*

Other notes on *p*-values:

1. Their use is controversial in some circles
2. Can be easily abused to show significant results
3. Despite limitations, they are ubiquitous in science
   - We have used them for so long, it's hard to change course (but Bayesians are trying!)
   - For many applied researchers and practitioners, they are a convenient way to turn observed data in to a "yes" / "no" decision

Deciding between H_0 and H_1

So, how do we choose between our hypotheses?

1. Our default is to believe H_0
2. We use our data to determine if we have sufficient reason to reject H_0
3. This is where we rely on work from probability theory

Deciding between H_0 and H_1 *cont.*

Because we are relying on probabilistic reasoning about whether or not to reject H_0, we can be wrong.

Our Decision	Truth: H_0 is True	Truth: H_0 is False
Reject H_0	Type I Error	Correct Decision
Fail to Reject H_0	Correct Decision	Type II Error

Deciding between H_0 and H_1 cont.

Question:

How do we know when we have committed a Type I error or a Type II error?

Deciding between H_0 and H_1 cont.

Answer:

In general, we cannot *know* unequivocally when we have committed a Type I error or a Type II error.

This has important implications:

1. Replication is *absolutely crucial* in science
2. Must be *hyper vigilant* about inflated Type I error from repeated testing (more on this later)
3. Should be generally skeptical, and especially so for low power studies with "sexy" results

The Binomial Test

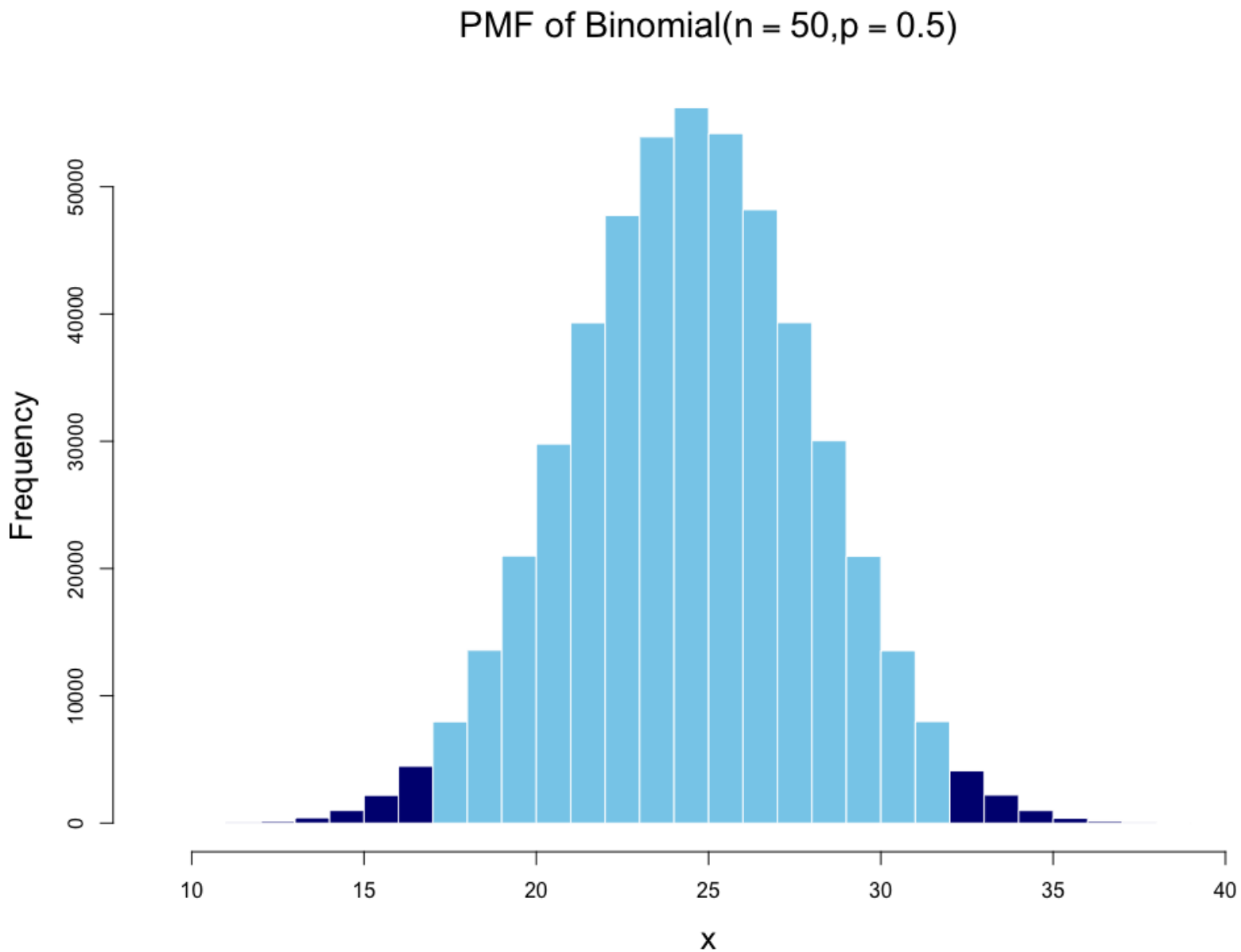
1. Probably the most basic example of a hypothesis tests (and very useful)
2. Used to compare distribution of observations in two categories against theoretical distribution
3. Essentially, we use the binomial test when we have a problem that can be expressed in terms of "successes" and "failures"

The Binomial Test *cont.*

Example questions we can answer:

1. Given N tosses of a coin, $X_1, X_2, ..., X_n$, where $X_i = 1$ and $X_i = 0$ is tails, is this a fair coin?
2. Given the counts of females and males in a particular class, are there significantly more females than males?
3. Suppose we are doing quality control on a medical device known to have a 0.001% failure rate. Given the number of failures in a specific batch and the batch size, does this batch have significantly more failures than we expect?

The Binomial Test *cont.*



Exact Binomial Test

Description:

Performs an exact test of a simple null hypothesis about the probability of success in a Bernoulli experiment.

Usage:

```
binom.test(x, n, p = 0.5,
           alternative = c("two.sided", "less", "greater"),
           conf.level = 0.95)
```

Arguments:

x: number of successes, or a vector of length 2 giving the numbers of successes and failures, respectively.

n: number of trials; ignored if ‘x’ has length 2.

p: hypothesized probability of success.

alternative: indicates the alternative hypothesis and must be one of ‘"two.sided"’, ‘"greater"’ or ‘"less"’. You can specify just the initial letter.

conf.level: confidence level for the returned confidence interval.

Binomial Test Example

<EXAMPLES_IN_R>

The Binomial Test *cont.*

What if your variable has more than 2 categories?

The Binomial Test *cont.*

There are a few options when you have a variable with more than 2 categories

1. Exact Multinomial Test (EMT package in R)
2. *G*-Test for Goodness-of-Fit (also called likelihood ratio test)
3. Pearson's χ^2 (Goodness-of-Fit) Test

Pearson's χ^2 (Goodness-of-Fit) Test

Pearson's χ^2 goodness-of-fit test can be used when we have some categorical variable, X, where each X_i is a value from one of K categories, and where $K \geq 2$ and we have an expected probability, P_k, for each category.

Pearson's χ^2 Test Example

Suppose we want to determine whether or not a die is loaded (i.e., not a fair die). Say we roll the die 100 times, and we obtain the following results:

Value	Count
1	13
2	21
3	15
4	17
5	20
6	14

Are we confident that this is a fair die?

Pearson's χ^2 Test Example *cont.*

The test statistic is χ^2 and is computed using:

$$\chi^2 = \sum_{k=1}^{K} \frac{(O_k - E_k)^2}{E_k},$$

where K is the number of categories, O_k is the observed count for category k, and E_k is the expected count for category k under the null hypothesis. The degrees of freedom are: $df = K - 1$.

Pearson's χ^2 Test Example *cont.*

The χ^2 test statistic follows the χ^2 distribution, a continuous distribution with a single parameter—the degrees of freedom (df).

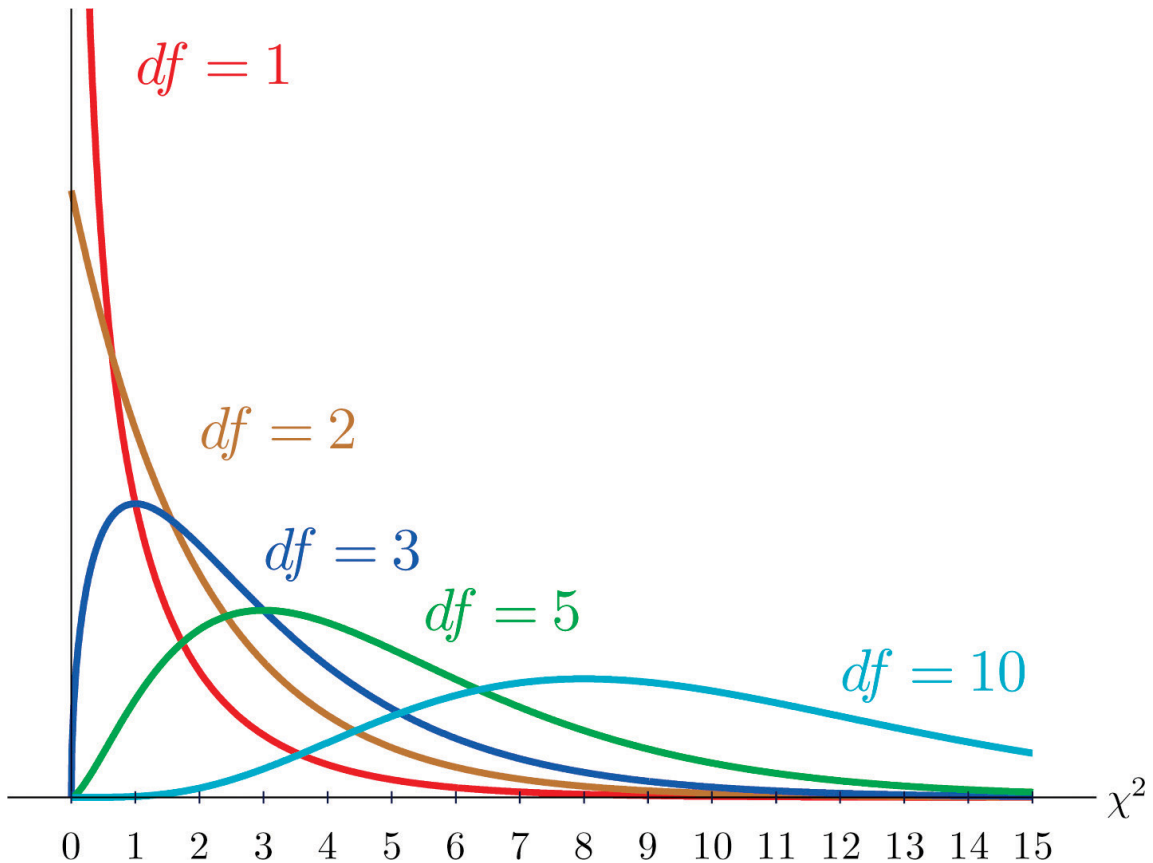


*Image source: https://stats.libretext.org

Pearson's χ^2 Test Example *cont.*

With this χ^2 and df, we evaluate probability of observed data if the null hypothesis is true.

- Note that Pearson's χ^2 goodness-of-fit test assumes observations are independent from one another

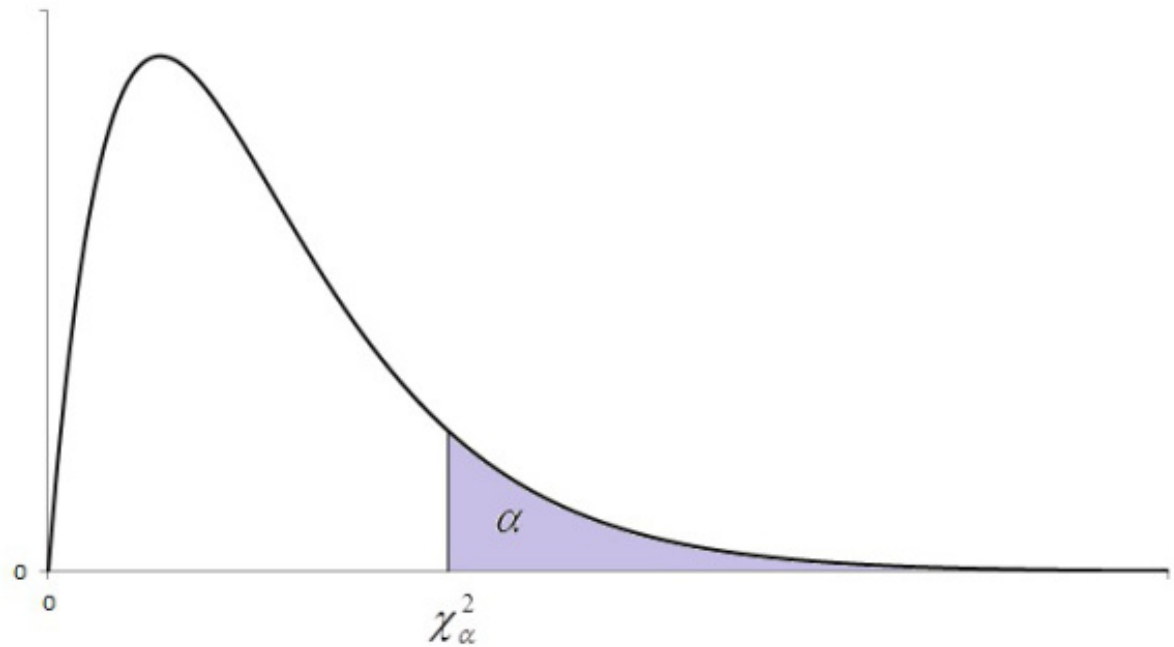


*Image source: https://actuarialmodelingtopics.wordpress.com

chisq.test package:stats R Documentation

Pearson's Chi-squared Test for Count Data

Description:

'chisq.test' performs chi-squared contingency table tests and goodness-of-fit tests.

Usage:

```
chisq.test(x, y = NULL, correct = TRUE,
           p = rep(1/length(x), length(x)), rescale.p = FALSE,
           simulate.p.value = FALSE, B = 2000)
```

Arguments:

x: a numeric vector or matrix. 'x' and 'y' can also both be factors.

y: a numeric vector; ignored if 'x' is a matrix. If 'x' is a factor, 'y' should be a factor of the same length.

Pearson's χ^2 (Goodness-of-Fit) Test Example

<EXAMPLES_IN_R>

Pearson's χ^2 (Independence) Test

We can also use Pearson's χ^2 to solve a different sort of problem. In particular, we can use Pearson's χ^2 to test the extent to which two categorical variables are independent.

Pearson's χ^2 (Independence) Test

Suppose we would like to teach cats to dance.

We have two training systems: using food as a reward, and using affection as a reward. Suppose after a week of training the cats, we test dancing ability. So, we have two categorical variables: *training* and *dance*, each with two levels.

Could cat dance?	Training Method: Food as Reward	Training Method: Affection as Reward
Yes	28	48
No	10	114

*Source: Field *et al.* 2012

Pearson's χ^2 Independence Test *cont.*

The test statistic is χ^2 and is computed using:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}},$$

where

$$E_{i,j} = \frac{\text{row-total}_i \times \text{column-total}_j}{N}$$

and where $O_{i,j}$ is the observed count in cell i, j and $E_{i,j}$ is the expected count for cell i, j under the null hypothesis.

Pearson's χ^2 Independence Test *cont.*

Note:

1. Degrees of freedom: $df = (r-1)(c-1)$ where r is the number of rows, and c is the number of columns
2. Assumption that observations are independent from one another
   - E.g., In above example, a cat could only be in one *training* condition

Pearson's χ^2 (Independence) Test Example

<EXAMPLES_IN_R>

Central Limit Theorem

Suppose that $X_1, X_2, ..., X_n$ is independently and identically distributed (*iid*) with mean μ and variance σ^2. According to the central limit theorem (CLT), $\overline{X}_n = (1/n)\sum_{i=1}^{n} X_i$ has a distribution that is approximately normal with mean μ and variance σ^2/n.

Central Limit Theorem *cont.*

This allows us to use $\sigma_{\overline{X}} = \frac{s}{\sqrt{n}}$, to approximate standard error (SE), where s is our sample standard deviation.

The importance of this will become clear as we discuss hypothesis testing with continuous variables.

Review of Standard Error

Recall that the term "standard error" (SE) refers to the standard deviation of the sampling distribution for a given parameter estimate.

And recall, the "sampling distribution" is the theoretical distribution of parameter estimates that would be obtained by collecting infinitely many samples

Confidence Intervals

Recall also that the SE plays a key role in constructing confidence interval. For example, the confidence interval for a mean (when we know the standard deviation) is:

$$\overline{x} \pm z^* \frac{\sigma}{\sqrt{n}},$$

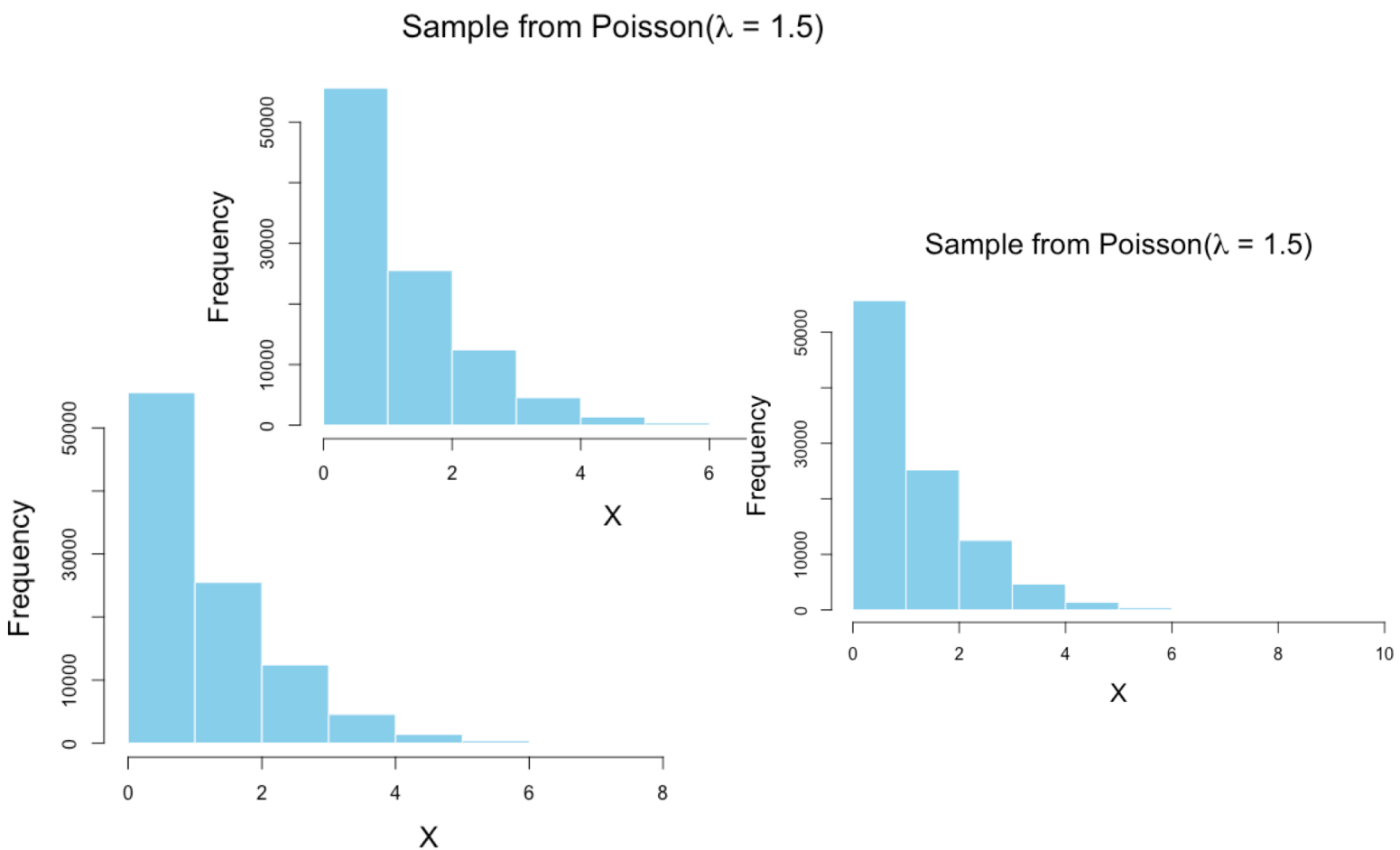
where z^* is the critical value from the normal distribution for the desired level of confidence (e.g., for 95% confidence, $z^* = 1.96$).

Central Limit Theorem *cont.*

The CLT also means that regardless of the original distribution of X, as long as its mean and variance exist, we can use the normal distribution to make probability statements about $\overline{X}_n$. This is true even in cases where the original distribution of X is discrete.

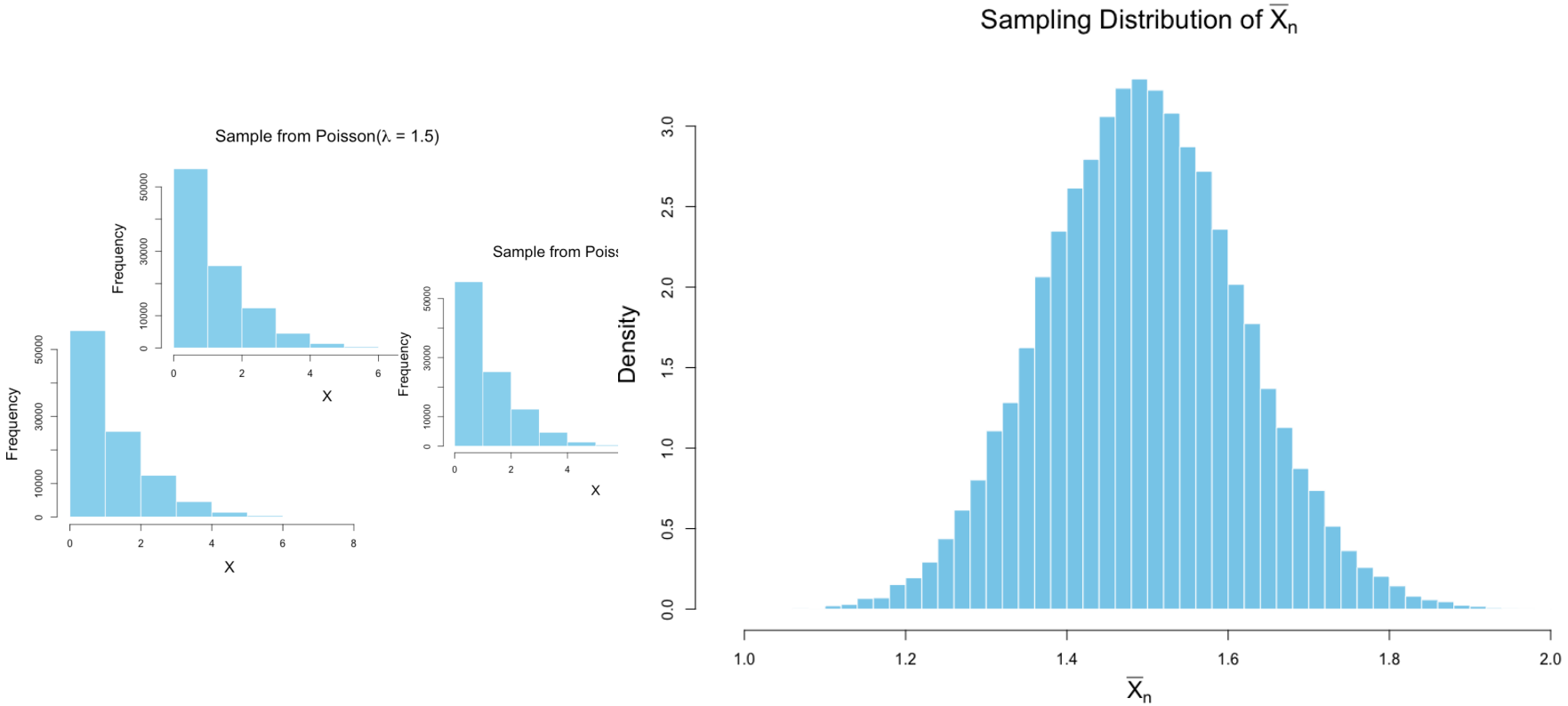
Central Limit Theorem *cont.*

Suppose we take samples of size n from a Poisson distribution with $\lambda = 1.5$



Central Limit Theorem *cont.*

Suppose we take samples of size n from a Poisson distribution with $\lambda = 1.5$



Continuous Data and Mean Comparisons

To this point, we have been looking at categorical data (e.g., "heads" / "tails", yes/no, cat dances/cat doesn't dance). Equipped with the CLT, we can now explore some interesting new methods. In particular, we can start looking at continuous variables.

Student's T-Test

The t-test refers to a family of statistical tests whose test statistic follows the t-distribution.

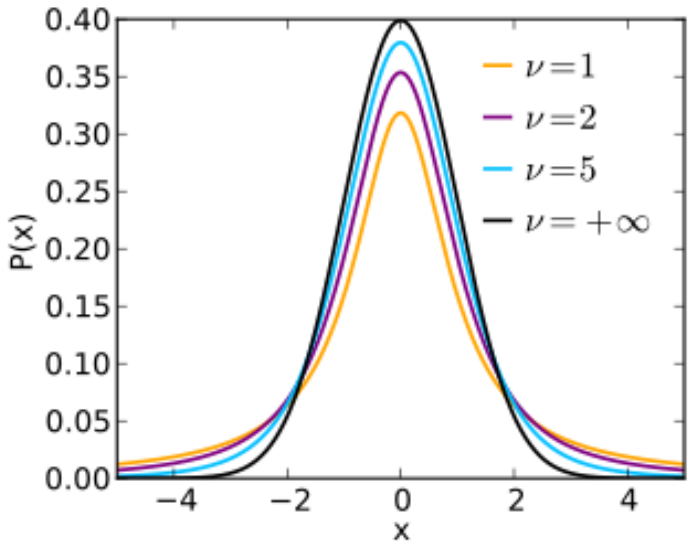


*Image source: wikipedia.org

Student's *T*-Test *cont.*

We will discuss three types of *t*-tests

1. One-sample *t*-test
2. Independent (two-sample) *t*-test
3. Dependent samples *t*-test
   - Also known as "paired-samples" *t*-test

General Notes on t-tests

1. All three versions of the t-test are implemented in R as the `t.test()` function.
   - Specifying different arguments to the function will give you different type of t-test.
2. In all three cases, the t-test can be done as one-sided or two sided. We will generally prefer two-sided tests.

One-Sample t-test

The one-sample t-test is used to test the null hypothesis that the population mean is equal to some value μ_0. The test statistics is defined as

$$t = \frac{\overline{x} - \mu_0}{\sigma_{\overline{x}}},$$

where $\overline{x}$ is the sample mean and $\sigma_{\overline{x}}$ is our estimate of the standard error of the mean. Recall it is defined as

$$\sigma_{\overline{x}} = \frac{s}{\sqrt{n}},$$

where s is the sample standard deviation and n is the sample size.

One-Sample t-test

So, our test statistics is defined as

$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}.$$

We also need to know the degrees of freedoms (ν) so we can compare our t to the appropriate t-distribution.

In the one-sample case, $\nu = n - 1$, where n is our sample size.

```
t.test                  package:stats                  R Documentation

Student's t-Test

Description:

     Performs one and two sample t-tests on vectors of data.

Usage:

     t.test(x, ...)

     ## Default S3 method:
     t.test(x, y = NULL,
            alternative = c("two.sided", "less", "greater"),
            mu = 0, paired = FALSE, var.equal = FALSE,
            conf.level = 0.95, ...)

     ## S3 method for class 'formula'
     t.test(formula, data, subset, na.action, ...)

Arguments:

       x: a (non-empty) numeric vector of data values.

       y: an optional (non-empty) numeric vector of data values.

alternative: a character string specifying the alternative hypothesis,
          must be one of '"two.sided"' (default), '"greater"' or
          '"less"'.  You can specify just the initial letter.

      mu: a number indicating the true value of the mean (or difference
          in means if you are performing a two sample test).

  paired: a logical indicating whether you want a paired t-test.
```

One-Sample *t*-test (*Example*)

Suppose you teach high school math and you'd like to know whether your students perform at, above, or below average on the math portion of the SAT.

<EXAMPLE_IN_R>

One-Sample *t*-test (*cont.*)

Assumptions of one-sample *t*-test:

1. Observations are independent
2. Variable is normally distributed in population
   - In practice, *t*-test is fairly robust to violations of normality provided n is not small.

Independent (Two-Sample) t-test

The independent t-test:

1. More common version of the t-test
2. Used to compare means from two different groups
3. Test statistic is:
$$t = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}},$$
where s_k^2 is variance of Group k, and n_k is sample size.
4. Our degrees of freedom are: $\nu = n_1 + n_2 - 2$

Independent (two-sample) *t*-test Example

Suppose we are interested in anxiety in arachnophobes. We want to know whether a photograph of a spider induces the same anxiety as an actual spider.

We recruit 24 arachnophobes, and show half of them a photograph of a spider, and the other half an actual spider. Both groups complete a survey measure of anxiety.

<EXAMPLE_IN_R>

*Source: Field *et al.* 2012

Independent (two-sample) *t*-test (*cont.*)

Assumptions of independent (two-sample) *t*-test:

1. Observations are independent
2. Variable is normally distributed in population
   - In practice, *t*-test is fairly robust to violations of normality provided n is not small.
3. Homogeneity of variance
   - Assume equal variances in two populations
   - The *t*-test is also quite fairly to violations of homogeneity of variance

Dependent (paired-sample) t-test

The paired t-test is often used when we have repeated measurements (i.e., one sample with two measurement occasions). The test statistics is defined as

$$t = \frac{\overline{x}_D - \mu_0}{\frac{s_D}{\sqrt{n}}},$$

Dependent t-test Example

Consider one of our first examples. Suppose we have developed some new medication to lower cholesterol. We randomly assign 50 patients each to a treatment and control group.

After 6 months, we measure their total cholesterol. We want to know if the treatment group's total cholesterol is different than the control group's.

<EXAMPLE_IN_R>

References

- Wasserman, L., (2004) "*All of Statistics*"
- Williams, R. (2017) "*Introduction to Hypothesis Testing*", https://www3.nd.edu/ rwilliam/stats1/x24.pdf