Andrew Gard - equitable.equations@gmail.com

What is the χ^2 Distribution?

Suppose you have a collection of independent numerical observations and would like to somehow measure how far they are from their expected values, in total. The variable χ^2 lets you do just this.

Suppose you have a collection of independent numerical observations and would like to somehow measure how far they are from their expected values, in total. The variable χ^2 lets you do just this. If the z-scores of the observations are $Z_1, Z_2, \ldots Z_r$, then we define

$$\chi^2 = Z_1^2 + Z_2^2 + \cdots + Z_r^2$$

Suppose you have a collection of independent numerical observations and would like to somehow measure how far they are from their expected values, in total. The variable χ^2 lets you do just this. If the z-scores of the observations are $Z_1, Z_2, \ldots Z_r$, then we define

$$\chi^2 = Z_1^2 + Z_2^2 + \cdots + Z_r^2$$

For instance, when all of the outcomes of the independent variables are exactly at their mean, χ^2 is exactly zero. As the individual results get more extreme, χ^2 gets larger. By squaring the z-scores before summing, we insure that low results don't cancel out high ones.

Suppose you have a collection of independent numerical observations and would like to somehow measure how far they are from their expected values, in total. The variable χ^2 lets you do just this. If the z-scores of the observations are $Z_1, Z_2, \ldots Z_r$, then we define

$$\chi^2 = Z_1^2 + Z_2^2 + \cdots + Z_r^2$$

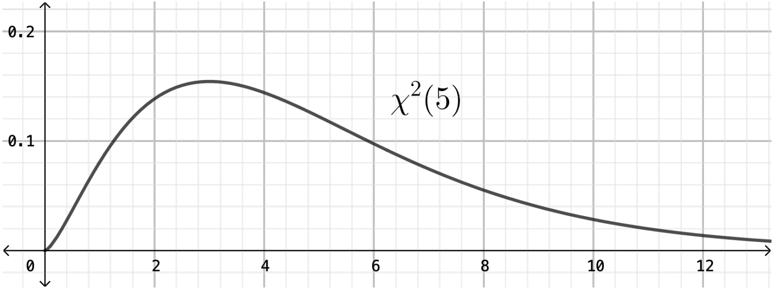
For instance, when all of the outcomes of the independent variables are exactly at their mean, χ^2 is exactly zero. As the individual results get more extreme, χ^2 gets larger. By squaring the z-scores before summing, we insure that low results don't cancel out high ones.

The sampling distribution of the random variable χ^2 is called the **χ^2-distribution with r degrees of freedom**, or $\chi^2(r)$ for short.

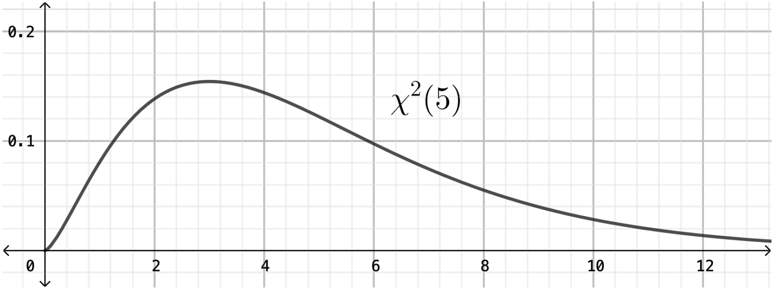
Since each Z_i is a continuous random variable, so is $\chi^2 = Z_1^2 + \cdots + Z_r^2$.

Since each Z_i is a continuous random variable, so is $\chi^2 = Z_1^2 + \cdots + Z_r^2$. The probability density function for χ^2 is zero for $\chi^2 \leq 0$ and is skewed to the right for any r.

Since each Z_i is a continuous random variable, so is $\chi^2 = Z_1^2 + \cdots + Z_r^2$. The probability density function for χ^2 is zero for $\chi^2 \leq 0$ and is skewed to the right for any r. The graph of $\chi^2(5)$ is typical.

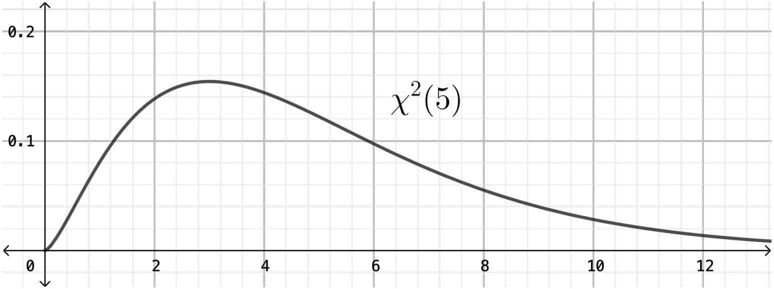


Since each Z_i is a continuous random variable, so is $\chi^2 = Z_1^2 + \cdots + Z_r^2$. The probability density function for χ^2 is zero for $\chi^2 \le 0$ and is skewed to the right for any r. The graph of $\chi^2(5)$ is typical.



Fact 1. The expected value of $\chi^2(r)$ is r.

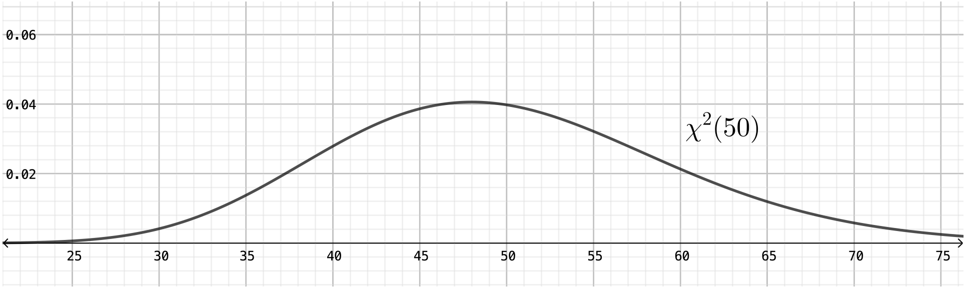
Since each Z_i is a continuous random variable, so is $\chi^2 = Z_1^2 + \cdots + Z_r^2$. The probability density function for χ^2 is zero for $\chi^2 \leq 0$ and is skewed to the right for any r. The graph of $\chi^2(5)$ is typical.



Fact 1. The expected value of $\chi^2(r)$ is r.

Fact 2. The mode (peak) of $\chi^2(r)$ is $r-2$ if $r \geq 2$ and zero otherwise.

When r is large, the distribution $\chi^2(r)$ is approximately normal. This is already visible when $r = 50$, as pictured below.



If you look closely you can still see the skew in this plot, however.

The χ^2 distribution comes up frequently in inferential statistics.

The χ^2 distribution comes up frequently in inferential statistics. A few of the most common applications are:

- **Significance testing for variance.** When drawing samples of size n from a normal distribution with hypothesized variance σ^2, the sampling distribution of $(n-1)S^2/\sigma^2$ is $\chi^2(n-1)$, where S^2 is the sample variance.

The χ^2 distribution comes up frequently in inferential statistics. A few of the most common applications are:

- **Significance testing for variance.** When drawing samples of size n from a normal distribution with hypothesized variance σ^2, the sampling distribution of $(n-1)S^2/\sigma^2$ is $\chi^2(n-1)$, where S^2 is the sample variance.
- **Goodness-of-fit testing.** When a categorical variable is hypothesized to have a certain distribution, the sampling distribution of $\sum \frac{(O-E)^2}{E}$ is approximately $\chi^2(n-1)$, where n is the number of categories, O is the observed count in each category, and E is the expected count under the hypothesized distribution.

The χ^2 distribution comes up frequently in inferential statistics. A few of the most common applications are:

- **Significance testing for variance.** When drawing samples of size n from a normal distribution with hypothesized variance σ^2, the sampling distribution of $(n-1)S^2/\sigma^2$ is $\chi^2(n-1)$, where S^2 is the sample variance.
- **Goodness-of-fit testing.** When a categorical variable is hypothesized to have a certain distribution, the sampling distribution of $\sum \frac{(O-E)^2}{E}$ is approximately $\chi^2(n-1)$, where n is the number of categories, O is the observed count in each category, and E is the expected count under the hypothesized distribution.
- **The χ^2-test for independence.** A similar test statistic can be used when testing whether two categorical variables are independent of one another.

As with any continuous random variable, we compute probabilities in χ^2 distributions using a cumulative distribution function, or cdf.

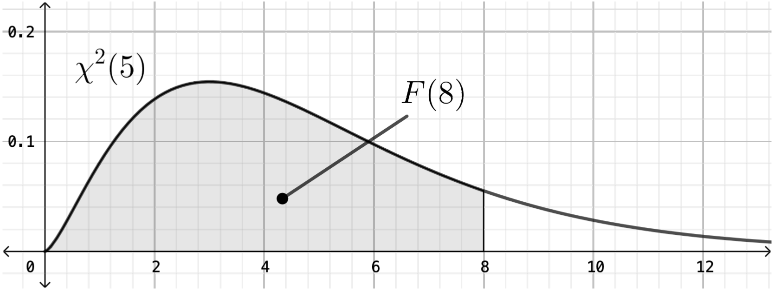
As with any continuous random variable, we compute probabilities in χ^2 distributions using a cumulative distribution function, or cdf. For any given value x, the cdf $F(x)$ gives the probability of randomly getting a value less than or equal to x in the appropriate χ^2 distribution. Technically,

$$F(x) = P(X \leq x) \;\; \text{in} \;\; \chi^2(r)$$

As with any continuous random variable, we compute probabilities in χ^2 distributions using a cumulative distribution function, or cdf. For any given value x, the cdf $F(x)$ gives the probability of randomly getting a value less than or equal to x in the appropriate χ^2 distribution. Technically,

$$F(x) = P(X \leq x) \text{ in } \chi^2(r)$$

For instance, $F(8) = P(X \leq 8)$ in $\chi^2(5)$ is pictured below.



In R, the cdf of $\chi^2(r)$ is represented by the command $pchisq(x, r)$, where x is the value of interest and r is the number of degrees of freedom.

In R, the cdf of $\chi^2(r)$ is represented by the command $pchisq(x, r)$, where x is the value of interest and r is the number of degrees of freedom. For instance, we can compute $F(8)$ in $\chi^2(5)$ using $pchisq(8, 5)$.

```r
pchisq(8, 5)
```

```
## [1] 0.8437644
```