STAT 206 - Statistics for Engineers

Kevin James

Fall 2013

Introduction

Statistics in the collection, organization, analysis, iterpretation, and presentation of data. In effect, it is a quantification of uncertainty.

Process

To conduct an **empirical study** or statistical study, we must first identify the **population** (the set of elements your query pertains to) of interest. Individual items of this population are called **units** (a single element, usually a person or object, whose characteristics we are interested in). We also define the hypothesis, or question we would like answered.

We select a subset of units from the population to have in our **sample** (a subset of the population from which measurements are actually made) which must have a pre-determined size and should make an attempt to reduce or eliminate **sample error** (an error which occurs randomly due to the uncertainty of the sample). We also must determine how we can measure the **variable of interest** (a measure of the interesting characteristic of a unit). This variable can often be measured in a multitude of ways, though many of which will be somewhat lacking in value. You must take into account not only what this variable is and how it is collected, but also ways to minimize bias, such as by randomizing and repeating your experiments.

We should also attempt to avoid **study errors** (systematic errors which occur because the sample does not accurately reflect the population) or else we will find ourselves with a large amount of error and/or uncertainty.

Post-experiments, we need to analyze our data and come to a conclusion. It is generally a goo didea to graph the data, as this gives us a highly visual method of analysis. We can use two main branches of statistics to analysyze our data: **descriptive statistics** (a summary of the collected data, both visuallt and numerically) or **inferential statistics** (generalized results for the population based on the sample data). We will be focusing on inferential statistics, which inlude a quantification of uncertainty, in this course.

Finally, we use the results of our study to answer the original hypothesis or research question. We also must be sure to address the limitations of our study.

Types of Variables

Our variables may be either **categorical** (a qualitative measure belonging to one of K possible classes), **discrete** (a quantitative measure with some countable value) or **continuous** (a quantitative measure with some uncountable value, such as a range of values).

Plots

We can design a **stem-and-leaf plot** by writing all first digits in a single column and all of the other digits in the corresponding right-hand side. For example, for a standard bell-curve grading scheme:

```
4 | 24
5 | 0068
6 | 24556
7 | 4556678889
8 | 00022223334558
9 | 0334469
```

We can also use **grouped frequency tables** by using frequency bins, for example

Average	Frequency
90+	18
80+	43
70+	87
60+	92

Histograms follow a similar pattern, since we select bins such as 40-49, 50-59, 60-69, 70-79, 80-89, 90-100 and diagram the amount in each bin. If we have differently sized bins (e.g. 1, 2, 3-4) we want to examine the "area" of the bars instead of their "height".

Measures of Certainty

The **sample mean** is a set of n values and is denoted

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

The **median** is the number x^* such that half the values are below and half are above. If we denote the i^{th} smallest value as x_i, then

$$x^* = x_{\frac{n+1}{2}}$$

if n is odd, or

$$x^* = \frac{x_{\frac{n+2}{2}} + x_{\frac{n}{2}}}{2}$$

if n is even.

Measures of Dispersion

The **sample variance** of a set of n values is denoted by

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}$$

The **standard deviation**, denoted s, is the square root of the sample variance.

The **range** of the set is the difference between the maximum and minimum values.

If we create a graph with the median, mean, average variance, etc, it is called a **box-and-whiskers** plot.

Probability

Classical probability is the "common sense" probability related to discrete events such as coin flips, dice rolls, etc. Though useful, this form of probability has some severe limitations: namely, the definition of what "equally likely" actually means. In effect, we can use this type of probability to find an answer, but can not use that answer for anything. **Relative frequency probability** is slightly more useful: we repeat an experiment some number of times and record the relative chance of various outcomes. This type of probability analysis, however, is extremely impractical. Finally, we have **subjective probability**, which is based on a person's experiences and subjective knowledge. Obviously, this method also has some severe limitations and is far too abstract to be used scientifically.

When discussing probabilty, we always refer to **experiments** (repeatable phenomena or processes) or their various **trials** (iterations of an experiment). These experiments have a **sample space** (set of discrete outcomes for an experiment). which is obviously either discrete or continuous, depending on whether or not this range is countable.

We will be attaching a mathematical model to the sample space to have our our definition of probability. Any probablity model must obey the following axioms:

- $0 \leq P(A) \leq 1, A \in S$
- $P(S) = 1$
- $P(A \cup B) = P(A) + P(B)$ for any mutually exclusive outcomes

for any sample space S and potential outcomes A and B.

The classical model would suggest that for a sample set $S = \{a, b, c\}$, each outcome has a probability $P = \frac{1}{3}$. This is referred to as a **uniform distribution**, and is incorrect for most non-trivial samples.

Permutations and Combinations

A common problem requires we create an arrangement using r of n objects. In such a set, the number of **permutations** is equal to

$$n^{(r)} = \frac{n!}{(n-r)!}$$

If we don't care about the order of the arrangement, we can use the formula for a **combination**. The way to find r of n items is

$$\binom{n}{r} = \frac{n!}{r!(n-r)!}$$

Set Operations

$A \cap B$ is the **intersection** of two events, event A and event B. In other words, this is the probability that both events will occur. It is also written as AB. Note that if $P(A \cap B) = 0$, the two events are mututally exclusive.

$P(A \cup B)$ is the **union** of events A and B, and is defined by $P(A \cup B) = P(A) + P(B) - P(\cap B)$. This is the probability of one or the other happening.

We also define the **complement** of A, $\overline{A} = 1 - P(A)$. This is the probability of the event not occuring.

We define **conditional occurances** with the following notation: the probability of A conditional on the probability of B is $P(A|B) = \frac{P(A \cap B)}{P(B)}$. Obviously, if the probability of B is zero, this is non-sensical.

Two events are **independant** if and only if $P(A \cap B) = P(A)P(B)$. Note that this will also tell us that $P(A|B) = P(A)$ and vice-versa (the probability of A/B is the same regardless of whether the other has occured.

Law of Total Probability

If we have some distinct partition of our sample set such that $B_0 \cup B_1 \cup \cdots B_n$, then for any event A we can find $P(A) = \sum_{i=0}^{n} P(A|B_i)P(i)$

Bayes' Theorem

For any two events in a sample set

$$P(B|A) = \frac{P(AB)}{P(A)} = \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|\overline{B})P(\overline{(}B)}$$

Discrete Random Variables

A **random variable** is one which may have any value, where $R(X)$ is the range of possible values it can take. We denote random variables with upper-case letters and denote observed variables as lower-case letters. If these variables can take on only two possible values, we refer to them as **binary**.

We denote the **probability distribution** (i.e. the chance of some random variable being equal to a certain variable) as $f(x) = P(X = x)$. The sum of the probability distributions of X for all possible x is equal to 1.

We also define the **cumulative distribution function** as $F(x) = P(X \leq x)$.

The **mean** or **expected value** of a random variable X is defined as

$$\mu = E(X) = \sum_x xf(x)$$

This function is linear, thus we have

$$E(aX + bY) = aE(X) + bE(Y)$$

Variance is the square of the expected difference

$$Var(X) = E((X - E(X))^2) = \sum_x f(x)(x - \mu)^2$$

We sometimes denote this as $Var(X) = E(X^2) - E(X)^2$.

Bernoulli Distributions

A Bernoulli distribution will be formed when an experiment is repeated several times. The outcomes of each tril must be independant though the probability of any given outcome must be identical over all experiments. Results must be binary.

We say that X follows a Bernoulli distribution ($X \sim Bernoulli(p)$), where p is the probablity of success, if

$$f(x) = \begin{cases} p & x = 1 \\ 1 - p & x = 0 \end{cases}$$

Note that for all Bernoulli distributions $E(X) = p$ and $Var(X) = p(1 - p)$.

Let X be the number of successes obtained from a sequence of n Bernoulli trials. X follows a **Binomial distribution** ($X \sim Bino(n; p)$) if

$$P(X = x) = f(x) = \binom{n}{x} p^x (1 - p)^{n-x}$$

We also have $E(X) = np$ and $Var(X) = np(1 - p)$.

When solving for a Bernoulli distribution, we may find that we give ourselves artificial boundaries. For example, if we have $n = 24$, we can not solve for $p > 24$. In this case, we can use limits to find the correct answer (i.e. $n = 24z$, $p = \frac{p_1}{z}$).

Binomial Theorem

For any positive integer n and real numbers a, b,

$$(a + b)^n = \sum_{x=0}^{n} \binom{n}{x} a^x b^{n-x}$$

Poisson Process

In a Poisson Process, events occur randomly in time or space according to the following conditions:

- **Independence** - the number of events in disjoint (i.e. non-overlapping) intervals are independant
- **Individuality** - events occur singlely (i.e. no two events can occur at a time)
- **Homogeneity** - events occur according to a uniform (constant) rate or intensity (λ)

If events occur with an average rate of λ per unit of time and X is the number of events which occur in t units of time, then $X \sim Poisson(\lambda t)$ gives us

$$f(x) = \frac{e^{-\lambda t}(\lambda t)^x}{x!}$$

for any $x = 0, 1, 2, \cdots$.

We can also define $E(X) = Var(X) = \lambda t$.

Hypergeometric Distribution

If we have a collection of N objects which can be sorted into two distinct types (success and failure), there exist r successes and $N - r$ failures. Assume we select a sample of n objects without replacements.

Then let X be the number of successes selected. X is said to follow a hypergeometric distribution $X \sim Hyper(N, r, n)$ if

$$f(x) = \frac{\binom{r}{x}\binom{N-r}{n-x}}{\binom{N}{n}}$$

for any $x = 0, 1, \cdots, \min(r, n)$. We can also define $E(X) = \frac{nr}{N}$ and $Var(X) = \frac{nr(N-r)(N-n)}{N^2(N-1)}$.

Geometric Distribution

In this case, Bernoulli trials are repeated until the first success. For X is the number of independant $Bernoulli(p)$ until the first success, then $X \sim Geometric(p)$ if

$$f(x) = p(1-p)^{x-1}$$

for any $x = 1, 2, 3, \cdots$

We also define $E(X) = \frac{1}{p}$ and $Var(X) = \frac{1-p}{p^2}$.

Continuous Probability

The sample space for continuous probability is all open intervals. In effect, the probability of an event occuring within a given interval is proportionate to the length/size of that interval (for uniform distributions).

Note that since ranges have probability (e.g. in total range 100, the sub-range $1 - 8$ has probability $\frac{8-1}{100} = .07$), individual elements must have 0 probability.

The **probablity density function** of a continuous random variable X describes the probability that X takes on a value in the range (a, b)

$$\int_a^b f(x)\ \mathrm{d}x = P(a < X < b)$$

The **cumulative density function** is the probability that $X < x$, or $F(x) = P(X < x)$.

Note that these two density functions are related in the following way:

$$\begin{aligned}\int_a^b f(x)\ \mathrm{d}x &= P(a < X < b) \\ &= P(X < b) - P(X < a) \\ &= F(b) - F(a)\end{aligned}$$

or in other words: $pdf(a, b) = cdf(b) - cdf(a)$.

Uniform Distribution

A continuous random variable is said to have a uniform distribution if the probability of a given subinterval is proportional to the length of that interval. If X is uniformly distributed over (a, b) we write $X \sim Unif(a, b)$, which gives us $f(x) = \frac{1}{b-a}$.

Normal Distribution

A random variable X has a normal distribution $X \sim N(\mu, \sigma^2)$ if the pdf takes the form

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$$

where the expected value is μ and the variance is σ^2.

If X has a normal distribution, then $Z = \frac{X-\mu}{\sigma} = N(0, 1)$. Also note that the normal distribution is symmetrical, i.e. $P(Z > z) = P(Z < -z)$.

To solve a normal distribution, we reduce it to $P(Z < x)$, where $x \in \mathbb{R}$, and look up the answer in a normal distribution table.

Binomial Distribution

For any $X \sim Binomial(n, p)$, if $np > 5$ and $n(1-p) > 5$ (i.e. n is large and p is near 0.5), we have

$$Z = \frac{X - np}{\sqrt{np(1-p)}} \sim N(0, 1)$$

Because we are using a continuous distribution to approximate a discrete distribution, we include a continuity correction:

$$P(a < X) \approx P\left(\frac{(a + .5) - np}{\sqrt{np(1-p)}} < Z\right)$$

$$P(X < b) \approx P\left(Z < \frac{(b - .5) - np}{\sqrt{np(1-p)}}\right)$$

Sampling

Statistical Interference

The goal of statistical inference is to draw conclusions about a population, given only a small sample of said population. We achieve a random subset of a population using **random sampling**.

Random Sampling

There are many common types of random sampling:

- For **Simple Random Sampling**, each unit in the population has the same chance of being selected.
- If we have distinct groups, **Stratified Random Sampling** may be an excellent option. .We first divide the population into K distinct samples; from these, we select varying amounts of random units:
  - We could select these through **equal allocation**, i.e. an equal number form each strata (simple random sampling).
  - We could use **proportional allocation**, i.e. random number of units propotional to the strata size.
  - We may try **Neyman (Optimal) Allocation**, where each sample is weighted by strata variance.
- For a low cost / more efficient solution, we may try **Cluster Sampling**. In this case, the population is divided into M natural clusters. We take a simple random sample of the clusters to get m clusters, and from these clusters we perform equal allocation.

Random Sample

A random sample of size n from an infinite population is a set of independant and identically distributed random variables. Each random variable has the same probability distribution, mean, and variance.

Central Limit Theorem

If we have a random sample, then for large values of X ($X > 25$), we have

$$Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$$

Confidence Intervals

For some random sample, if the probability funciton of X_i depends on some unknown parameter θ, then we say an estimator of θ is a function of the sample

$$\hat{\theta} = h(X_1, X_2, ...X_n)$$

An **estimator** is said to be unbiased if the expected value is itself, i.e. $E(\hat{\theta}) = \theta$. The standard deviation, or standard error, of an esitmator is equal to $SE(\hat{\theta}) = \sqrt{Var(\theta)}$. Note that if we have two estimators for a parameter, the one with the lower standard error will be more efficient.

A $(1-\alpha)\%$ **confidence interval** for a parameter θ is an observation of the random interval $(L(X), U(X))$ such that

$$P(L(X) < \theta < U(X)) = (1 - \alpha)$$

Note that in this case it is the ends of the interval $L(X)$ and $U(X)$ which are random, not θ itself.

This also gives us α probability that the random interval does not contain the true value of the parameter.

Consider $X \sim N(\mu, \sigma^2)$, where σ^2 is a known value. A $(1 - \alpha)\%$ confidence interval for μ is given by

$$\overline{x} \pm z_{1-\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}}$$

where z_α is the α-quantile, i.e. the value such that $P(Z < z_\alpha) = \alpha$.

The **margin of error** of a confidence interval is the distance between the point estimate and the ends of the interval. For the above confidene interval, the margin of error is given by $z_{1-\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}}$.

Binomial Confidence

For a single $X \sim Binomial$, the probability of success can be estimated as

$$\hat{p} = \frac{X}{n}$$

Using Binomial approximation, we have

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0, 1)$$

An approximate $(1 - \alpha)\%$ confidence interval for p is given by

$$\hat{p} \pm z_{1-\frac{\alpha}{2}} \times \frac{\hat{p}(1 - \hat{p}}{\sqrt{n}}$$

χ^2 Distribution

If $X_1, X_2, ...X_n$ are in the normal distribution and $S = \sum_{i=1}^{n} X_i^2$ then we have a **chi-squared distibution** with n degrees of freedom

$$S \sim \chi^2[n]$$

The pdf of a chi-squared distribution is not integrable, so we must use the χ^2 table to calculate probabilities.

For a χ^2 distribution S, we have $E(S) = n$, $Var(S) = 2n$. For two independent χ^2 distributions S_1 and S_2, $S_1 + S_2 \sim \chi^2[n_1 + n_2]$.

If $X_1, X_2, \ldots X_n \sim N(\mu, \sigma^2)$ and their sample variance is $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ then we have

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2[n-1]$$

and furthermore S^2 is independant of $\overline{X}$.

Normal CI for σ^2

Consider $X_i \sim N(\mu, \sigma^2)$. We have the pivot quantity

$$S = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2[n-1]$$

and the $(1-\alpha)\%$ confidence interval for σ^2 given by

$$\left(\frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2}}[n-1]}, \frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2}}[n-1]}\right)$$

where $\chi^2_\alpha[n-1]$ is a value such that $P(S < \chi^2_\alpha[n-1]) = \alpha$.

Confidence Intervals for Two Means

Besides CIs for a single mean, we can also calculate the CI of multiple means: either the exact CI for paired means or an approximate CI for independant ones.

Paired Means

Suppose we have paired observations (x_i, y_i), where $X \sim N(\mu_X, \sigma^2_X)$ and $Y \sim N(\mu_Y, \sigma^2_Y)$ are two random variables measured from the same unit of the population. If we define $D = X - Y$ with observations $d_i = x_i = y_i$ then a $(1-\alpha)\%$ CI for μ_D using single means methods is given by

$$\overline{D} \pm t_{1-\frac{\alpha}{2}}[n-1]\frac{S_D}{\sqrt{n}}$$

The same CI can be given for $\mu_X - \mu_Y$, of course.

Independant Observations

Suppose X_i and Y_i are independent samples of lengths n_X and n_Y. For large length values, we have the pivotal quantity

$$\frac{(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}} \sim N(0,1)$$

An approximate CI is given by

$$(\overline{x} - \overline{y}) \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}}$$

Non-Symmetric Confidence Intervals

We can find the upper margin of error with

$$\frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2}}[n-1]} - S^2$$

or the lower with

$$S^2 - \frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2}}[n-1]}$$

The difference between the upper and the lower margins of error, of course, gives us the average margin of error.

Hypothesis Testing

The null hypothesis H_0 is a statement about the population or parameter of interest. The alternative hypothesis H_a is the conclusion we support if and only if there is sufficient evidence to reject H_0.

We define the **test statistic** as the evidence agains the null hypothesis. A value of zero indicates complete agreement; larger values indicate more evidence against. We reject the hypothesis based on the comparision with some critical value (i.e. if $t_{obs} > t_{crit}$, we reject the null hypohesis).

The are two types of **errors** involved in hypothesis testing

- Type I Error: Reject the null hypotheses when it is actually true. The probability of this error is α
- Type II Error: Do not reject the null hypothesis though it is actually false. The probability of this error is β

where α is the **significance level** of the test.

The **p-value** $\hat{p}$ of a test is the probability that the test statistic is greater than its observed value. By the range within which the p-value falls, we can determine how much eviedence exists against the null hypothesis.

- $\hat{p} > 0.1$: No evidence against the null hypothesis.
- $0.05 < \hat{p} < 0.1$: Some evidence against the null hypothesis.

- $0.01 < \hat{p} < 0.05$: Evidence against the null hypothesis.
- $\hat{p} < 0.01$: Strong evidence against the null hypothesis.

There is a relationship between p-values and test statistics

$$t_{obs} > t_{crit} \iff \hat{p} < \alpha$$

As such, the p-value may be used directly to determine the result of a hypothesis test.

The **power** of a hypothesis test is the probability that it rejects the null hypothesis AND the null hypothesis is actually false.

$$\text{power} = 1 - \beta$$

Note that we can have either one- or two-sided hypotheses, e.g. we could test $p = 0.5$ against $p \neq 0.5$ or we could test $p > 5$ against $p < 5$.

Independence Testing

When we are attempting to test the independence of two samples, we have

$$\chi^2_{obs} = \sum_{i,j} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}$$

where $O_{i,j}$ is the observed value of each variable for each sample and $E_{i,j}$ is its expected value.

Linear Regression

To **regress** is to search for a relationship between paired random variables. **Simple Linear Regression** is to quantify a linear relationship between them.

We define the following:

$$S_{xx} = \sum_{i=1}^{n} (x_i - \overline{x})^2$$
$$S_{xy} = \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})$$
$$S_{yy} = \sum_{i=1}^{n} (y_i - \overline{y})^2$$

The **correlation coefficient** between two random samples is given as

$$\rho = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$$

Correlation measures the strength of the relationship between the two variables; a value of ± 1 corresponds to an exact relationship, a value of 0 corresponds to no relationship.

Least squares estimators seek to minimize the sum of the squares of residuals $\varepsilon_i = y_i - a - bx_i$ with

$$L = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

To get the least square estimates, we find $\frac{\mathrm{d}L}{\mathrm{d}a} = 0$ and $\frac{\mathrm{d}L}{\mathrm{d}b} = 0$.

The estimates are given by $\hat{b} = \frac{S_{xy}}{S_{xx}}$ and $\hat{a} = \overline{y} - \hat{b}\overline{b}$.

Summary of Distributions

Bernoulli

Repeated trials of either success or failure (i.e. binary). The probability of success is the same for each trial; outcomes are independant.

For ditribution X with some individual trial x, given the probability of x is p:

$$\begin{aligned} X\sim\text{Bernoulli}(p) &\implies f(x) = p \\ &\implies E(X) = p \\ &\implies \text{Var}(X) = p(1-p) \end{aligned}$$

Binomial

A sequence of n independent Bernoulli trials.

For distribution X with n trials x each with probability p:

$$\begin{aligned} X\sim\text{Bino}(n,p) &\implies f(x) = \binom{n}{x} p^x (1-p)^{n-x} \\ &\implies E(X) = np \\ &\implies \text{Var}(X) = np(1-p) \end{aligned}$$

Poisson

Events occuring in time: they must be independent, non-simultaneous, and have a constant rate.

For distribution X with events x occuring at a rate of λ over a time interval t:

$$\begin{aligned} X\sim\text{Poisson}(\lambda,t) &\implies f(x) = \frac{e^{-\lambda t}(\lambda t)^x}{x!} \\ &\implies E(X) = \lambda t \\ &\implies \text{Var}(X) = \lambda t \end{aligned}$$

Hypergeometric

For a collection of N binary objects with r successes where n values are selected without replacement:

$$\begin{aligned} X\sim\text{Hyper}(N,r,n) &\implies f(x) = \frac{\binom{r}{x}\binom{N-r}{n-x}}{\binom{N}{n}} \\ &\implies E(X) = \frac{nr}{N} \\ &\implies \text{Var}(X) = \frac{nr(N-r)(N-n)}{N^2(N-1)} \end{aligned}$$

Geometric

A sequence of Bernoulli trials repeated until the first success.

$$\begin{aligned} X\sim\text{Geometric}(p) &\implies f(x) = p(1-p)^{x-1} \\ &\implies E(X) = \frac{1}{p} \\ &\implies \text{Var}(X) = \frac{1-p}{p^2} \end{aligned}$$

Continuous Distributions

The **cdf** is the probablity that $\overline{X}$ (a continuous random variable) is less than x, i.e.

$$\text{cdf}(y) = P(X < x)$$

The cdf is the integral of the **pdf**, i.e.

$$\text{cdf}(x) = \int \text{pdf}(x)$$

Any continuous random variable has

$$E(X) = \int_x f(x)\ \mathrm{d}x = \mu$$

and

$$\text{Var}(x) = E(x^2) - (E(x))^2$$

Uniform Distribution

The probability of any subinterval on the range is proportional to its length.

For a distribution X on an interval (a, b):

$$\begin{aligned} X\sim\text{Unif}(a,b) &\implies f(x) = \frac{1}{b-a} \\ &\implies E(X) = \frac{a+b}{2} \\ &\implies \text{Var}(X) = \frac{(b-a)^2}{2} \end{aligned}$$

Exponential Distribution

Events occur according to a Poisson process and we are measuring the time between events.

For X is the time until the next event and θ is $\frac{1}{\lambda}$:

$$\begin{aligned} X{\sim}\text{Exp}(\theta) &\implies f(x) = \frac{1}{\theta}e^{-\frac{x}{\theta}} \\ &\implies E(X) = \theta \\ &\implies \text{Var}(X) = \theta^2 \end{aligned}$$

Normal Distributions

For a distribution X with total average μ and total variance σ^2:

$$\begin{aligned} X{\sim}\text{N}(\mu, \sigma^2) &\implies f(x) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} \\ &\implies E(X) = \mu \\ &\implies \text{Var}(X) = \sigma^2 \end{aligned}$$

Binomial Approximations

For a Binomial distribution $X{\sim}\text{Binomial}(n, p)$ with $np > 5$, we have

$$Z = \frac{X - np}{\sqrt{np(1-p)}}{\sim}\text{N}(0, 1)$$

Since a Normal distribution is continuous and a Binomial distribution is discrete, we have to correct this with

$$\begin{aligned} P(a < X) &\approx P\left(\frac{(a - 0.5) - np}{\sqrt{np(1-p)}} < Z\right) \\ P(X < b) &\approx P\left(Z < \frac{(b + 0.5) - np}{\sqrt{np(1-p)}}\right) \end{aligned}$$

Sampling

When sampling, we use $\overline{X}$ to be the mean of the sample $\frac{1}{n}\sum_1 nx_i$ and S^2 to be the variance of the sample $\frac{1}{n-1}\sum_1 n(x_i - \overline{X})^2$. Note that this is different than the mean of the population σ^2. For a random sample, $E(\overline{X}) = \mu$ and $\text{Var}(\overline{X}) = \frac{\sigma^2}{n}$.

Normal Sample

A sample from a population with a normal distribution is represented as

$$X{\sim}\text{N}(\mu, \sigma^2) \implies \overline{X}{\sim}\text{N}(\mu, \frac{\sigma^2}{n})$$

Central Limit Theorem

If the sample is a normal distribution, it is convenient to give it a new pivotal point at zero with a variance of one (so we can look the probabilities up in a chart).

For the average of the observations $\overline{X}$, the actual average μ, the actual variance σ^2, and the number of observations n, we have

$$Z = \frac{\overline{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \sim \mathrm{N}(0,1)$$

Note that this is usually only valid for $n > 25$.

Confidence Intervals

A confidence interval is the interval on which we are $\alpha\%$ sure an oservation will fail.

For a normal distribution $\mathrm{N}(\mu, \sigma^2)$ with σ^2 known, we have the margin of error

$$\overline{X} \pm Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma^2}{n}}$$

Binomial CI

For a confidence interval with probability of success $\hat{p} = \frac{x}{n}$ we have

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim \mathrm{N}(0,1)$$

where we can find the margin of error as

$$\hat{p} \pm Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

χ^2 (Chi-squared) CI

For finding the CI of a variance, we have the sum of observations squared $S \sim \chi^2[n]$, where $E(S) = n$ and $\mathrm{Var}(S) = 2n$. If any two sums S_1 and S_2 are independent, we have

$$S_1 + S_2 \sim \chi^2[n_1 + n_2]$$

Since the variance S^2 can be found by

$$S^2 = \frac{1}{n-1} \sum (x_i - \overline{X})^2$$

then we have our pivotal quantity

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2[n-1]$$

Thus our CI is given as

$$\left(\frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2}}[n-1]} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2}}[n-1]} \right)$$

t Distribution

For a mean with an unknown variance, with $Z \sim \mathrm{N}(0,1)$ independent of $S \sim \chi^2[n]$, we have

$$\begin{aligned} T &= \frac{Z}{\sqrt{\frac{S^2}{n}}} \sim t[n] \\ &= \frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}} \sim t[n-1] \end{aligned}$$

Comparing Populations

If both populations are normal and dependant, $D = X - Y \sim \mathrm{N}(\mu_0, \sigma_0^2)$, where each trial is given by $d_i = x_i = y_i$. The CI for the mean is

$$\overline{d} \pm t_{1-\frac{\alpha}{2}}[n-1]\frac{Sd}{\sqrt{n}}$$

where we have $\overline{d} = \overline{x} - \overline{y}$.

If X and Y are measured over different samples, we have $E(X) = \mu_x$ and $E(Y) = \mu_y$, $\mathrm{Var}(X) = \sigma_x^2$ and $\mathrm{Var}(Y) = \sigma_y^2$, and

$$Z = \frac{(\overline{x} - \overline{y}) - (\mu_x - \mu_y)}{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}} \sim \mathrm{N}(0,1)$$

The CI can then be found with

$$(\overline{X} - \overline{Y}) \pm Z_{1-\frac{\alpha}{2}}\sqrt{\frac{S_x^2}{n_x} + \frac{S_y^2}{n_y}}$$