A Tour of Probability and Statistics for Math 4/896, Section 006

Thomas Shores
Department of Mathematics
University of Nebraska
Spring 2006

Contents

Note: This really is a *brief* tour. You will find much more detail in the excellent probability and statistics review in Appendix B of our textbook. Everyone should read through this appendix.

Probability

We'll begin with a few simple examples, one with a discrete set of outcomes and the other a continuous set.

Example 0.1. Consider the experiment of randomly selecting an individual out of the entire population of a certain species of animal for the purpose of some measurement. The selection of a particular individual could be thought of as an *outcome* to this random experiment. Selection of a male would amount to an *event* E , and the probability of selecting a male would be a number $P(E)$ between 0 and 1.

Example 0.2. Consider the experiment of throwing a dart at a dart board. We assume that the throw always hits the dart board somewhere. Here the outcome of this experiment is to locate the dart on some point on the dart board, so we can think of these points as outcomes. One event of interest is the event E of hitting the bulls eye region with the dart. Again, the probability of doing so would be a number $P(E)$ between 0 and 1.

Here are some of the key concepts of probability theory. You should relate these to the two experiments just described.

- **Sample Space:** A set S of possible outcomes from a random experiment or sequence thereof.
- **Event:** Any subset of the sample space S, i.e., of outcomes. (In some cases, there might be limitations on what subsets are admissible.)
- **Probability measure:** A way of measuring the likelihood that an outcome belongs to event E. This is a function $P(E)$ of events with natural properties: $0 \le P(E) \le 1$, $P(S) = 1$ and for disjoint events E_i,

  $$\sum_{i=1}^{\infty} P(E_i) = P\left(\bigcup_{i=1}^{\infty} E_i\right).$$

  Simple_consequence: If E is an event, then the probability of the complementary event $\overline{E}$ occurring is

  $$P\left(\overline{E}\right) = 1 - P(E)$$

- **Conditional probability:** This is the probability that an event E occurs, given that event F has occurred. It is denoted and defined by the formula

  $$P(E \,|\, F) = \frac{P(EF)}{P(F)}.$$

  Note the notation EF, which means the event of the occurrence of *both* E and F. Another way of expressing this event is the set-theoretic notation $E \cap F$.
- **Independent events:** Events E and F such that

  $$P(EF) = P(E)\,P(F)$$

  in which case the conditional probability of E given F is

  $$P(E \mid F) \equiv \frac{P(EF)}{P(F)} = P(E).$$

- **Law of Total Probability:** Given disjoint and exhaustive events $E_1, E_{2,}, \ldots, E_n$, and another event F,

  $$P(F) = \sum_{i=1}^{n} P(F \,|\, E_i)\, P(E_j)$$

- **Bayes' Theorem:**

  $$P(E \mid F) \equiv \frac{P(F \mid E)\, P(E)}{P(F)}.$$

  Some writers identify Bayes' Theorem as a combination of the Law of Total Probability and the above, namely, with notation as in the LTP and index k,

  $$P(E_k \mid F) \equiv \frac{P(F \mid E_k)\, P(E_k)}{\sum_{i=1}^{n} P(F \,|\, E_i)\, P(E_j)}.$$

Random Variables. Once we have randomly selected an individual outcome ω in an experiment, we can observe some relevant quantity and call it $X(\omega)$. This function X is called a **random variable**, *and a particular value observed in an experiment is customarily denoted as* $x = X(\omega)$.

Let's review the standard notations of this statistical framework.

- **Random variable:** a function X (abbreviate to r.v.) mapping outcomes to real numbers. A particular value is denoted by lower case x.
- **Probability density function:** a function p mapping the range of a random variable to probabilities (abbreviate to p.d.f.):
  In the case the r.v. is *discrete*, say has values $x_1, x_2, \ldots$ then
  $$P(a \le X \le b) = \sum \{p(x_i) \mid a \le x_i \le b\}.$$
  In the discrete case, $p(x)$ is also referred to as a *probability mass function* (p.m.f.). If the r.v. is continuous, then the density function f satisfies
  $$P(a \le X \le b) = \int_a^b f(x)dx.$$
  In this case f really is a density function with units of of probability per length. Note: since an experiment always results in *some* value of the r.v. X, we must have $\int_{-\infty}^{\infty} f(x)\, dx = 1$ and a similar result for discrete r.v.'s
- The **(cumulative) distribution function** (abbreviate to c.d.f.) associated to the r.v. is
  $$p(x) = P(X \le x) = \sum \{p(x_i) \mid x_i \le x\}$$
  for discrete r.v.'s and
  $$F(x) = P(X \le x) = \int_{-\infty}^{x} f(s)ds$$
  for continuous r.v.'s. Note: from properties of the p.d.f., we see that
  - $F(x)$ is a monotone increasing function, i.e., if $x \le y$, then $F(x) \le F(y)$.
  - $\lim_{x \to \infty} F(x) = 1$.

Discrete.

Example 0.3. Consider the experiment of Example_0.2. Once we have thrown the dart and landed on ω, we might observe the score $X(\omega)$ we earned according to the portion of the dart board on which our dart landed. Here X will take on a finite number of values. Let us further suppose that there are only two areas on the board: the center bullseye of area A (winner, value 1) and an outer area B (loser, value 0.) Suppose that the probability of hitting one area is proportional to its area. Then the probability of hitting the bullseye is

$$p = \frac{A}{A+B}$$

and the probability of losing is $q = 1 - p$. The p.d.f. is given by $f(0) = q$ and $f(1) = p$. The c.d.f. is given by $F(0) = q$ and $f(1) = 1$.

An interesting variation on the previous example is to repeat the experiment, say n times. Now the random variable X is your score: the number of times you hit the bullseye. The p.d.f. for this experiment (called a Bernoulli trial) is the so-called **binomial distribution**

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}, \; x = 0, 1, \ldots, n$$

Continuous.

Example 0.4. Consider the experiment of Example 0.1. Once we have selected an animal ω, we might take its weight and call it the statistic $X(\omega)$. Note that X could take on a continuous range of values. The p.d.f. and c.d.f. of a continuous random variable are more subtle and one often makes a priori assumptions about them.

Let's simplify our dart example, so we can obtain distributions more easily.

Example 0.5. Suppose that our target is not a two dimensional board, but a one dimensional line segment, say the interval of points x such that $a \le x \le b$ or symbolically, $[a, b]$ Suppose further that there is no bias toward any one point. Then it is reasonable to assume that the p.d.f. is constant. Since it is defined on $[a, b]$ and the area under this function should be 1, we see that the p.d.f. is the function

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{if } a \le x \le b \\ 0 & \text{otherwise} \end{cases}$$

while the c.d.f. should be

$$F(x) = \begin{cases} 0 & \text{if } x < a \\ \frac{1}{b-a}(x-a) & \text{if } a \le x \le b \\ 1 & \text{if } b < x. \end{cases}$$

This is the so-called **uniform distribution**.

Before we discuss further specific distributions, there are some more concepts we should develop.

Expectation and Variance. Key concepts:

- **Expectation** of a function g of a r.v.:

$$E[g(X)] = \begin{cases} \sum_i g(x_i) p(x_i), & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} g(x) f(x) dx & \text{if } X \text{ is continuous} \end{cases}$$

- **Expectation** of X (or mean, first moment): $\mu = \mu_X = E[X]$. One can show

$$\begin{aligned} E[\alpha X + \beta] &= \alpha E[X] + \beta \\ E[\alpha X + \beta Y] &= \alpha E[X] + \beta E[Y] \end{aligned}$$

- **Variance** of X: This is just

$$\mathrm{Var}(X) = E\left[(X - E[X])^2\right].$$

One can show

$$\begin{aligned} \mathrm{Var}(X) &= E\left[X^2\right] - E[X]^2 \\ \mathrm{Var}(\alpha X + \beta) &= \alpha^2 \, \mathrm{Var}(X) \end{aligned}$$

- **Standard deviation** of X: $\sigma = \sigma_X = \operatorname{Var}(X)^{1/2}$

Basically, the idea is this: the expected value is a kind of weighted average of values, so that one could say roughly that "on the average one expects the value of repeated experiments to be the mean." The variance and standard deviation are measures of the spread of the random variable. Note that the units of σ are the same as the units of μ, so that the standard deviation is a more practical measure of the spread of the random variable, but the variance σ^2 is more useful for some calculations and theoretical purposes.

Standard Notation: To save ourselves the inconvenience of always having to assign a name to the p.d.f. and c.d.f. of a given r.v. X, we adopt the convention that

$$\begin{aligned} f_X(x) &= \text{p.d.f. of the r.v. } X \\ F_X(x) &= \text{c.d.f. of the r.v. } X. \end{aligned}$$

Normality and the Central Limit Theorem. One of the most important single distributions in statistics is the **normal distribution**. This is a r.v. whose density function is the famous bell shaped curve

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}, \ -\infty < x < \infty.$$

It can be shown that this really is a density function with mean μ and variance σ^2. Its corresponding distribution is

$$F(x) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{x} e^{-(s-\mu)^2/2\sigma^2} ds, \ -\infty < x < \infty.$$

The **standard normal distribution** is the one with $\mu = 0$ and $\sigma = 1$, that is,

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \ -\infty < x < \infty$$

is the p.d.f. of the distribution. The c.d.f. for the standard normal distribution has the following designation, which we use throughout our discussion of statistics:

$$N(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-s^2/2} ds, \ -\infty < x < \infty.$$

The notation $N(\mu, \sigma^2)$ is used for a normal distribution of mean μ and variance σ^2. One sees phrases like "X is $N(\mu, \sigma^2)$" or "$X \ \ N(\mu, \sigma^2)$." We can pass back and forth between standard normal distributions because of this important fact: if X has a distribution $N(\mu, \sigma^2)$, then $Z = (X - \mu)/\sigma$ has the distribution $N(0, 1)$, the standard normal distribution.

Here is a key property of this important kind of distribution:

Theorem: If X and Y are independent normal random variables with parameters (μ_1, σ_1^2), (μ_2, σ_2^2), then $X + Y$ is normal with parameters $(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.

It follows that this Theorem is true for any finite number of independent r.v.'s.

In a limiting sense, sums of r.v.'s with finite mean and variance tend to a normal distribution. This is the Central Limit Theorem.

Central Limit Theorem. Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed random variables with a finite expected value μ and variance σ^2. Then the random variable

$$Z_n = \frac{X_1 + X_2 + \cdots + X_n - n\mu}{\sqrt{n}\sigma} = \frac{\frac{1}{n}(X_1 + X_2 + \cdots + X_n) - \mu}{\sigma/\sqrt{n}}$$

has distribution that approaches the standard normal distribution as $n \to \infty$.

Some Common Distributions. Here are a few common distributions.

Binomial:

- $f(x) = \dfrac{n!}{x!\,(n-x)!} p^x (1-p)^{n-x}$, $x = 0, 1, \ldots, n$
- Mean: $\mu = np$
- Variance: $\sigma^2 = np(1-p)$
- Application: Bernoulli trials as in variation on Example 0.4

Poisson:

- $f(x) = \dfrac{\mu^x e^{-\mu}}{x!} p^x (1-p)^{n-x}$, $n = 0, 1, \ldots$
- Mean: $\mu = \mu$
- Variance: $\sigma^2 = \mu$
- Application: A limiting case of binomial distribution. Used, e.g., to approximate binomial distributions with large n and $\mu = np$ of moderate size (typically < 5.) There is a whole family of "Poisson processes" that are used in problems like manufacturing errors, etc.

Gamma:

- $f(x) = \dfrac{1}{\Gamma(\alpha)\,\beta^\alpha} x^{\alpha-1} e^{-x/\beta}$, $0 < x < \infty$, $\Gamma(x) = \int_0^\infty s^{x-1} e^{-s}\, ds$. Here $2\alpha = \nu$ is called the number of *degrees of freedom.*
- Mean: $\mu = \alpha\beta$
- Variance: $\sigma^2 = \alpha\beta^2$
- Application: An umbrella for other extremely important p.d.f.'s. For example, $\alpha = 1$, $\beta = 1/\lambda$ gives the family of exponential distributions and $\alpha = \nu/2$, $\beta = 2$ gives a chi-square distribution with ν degrees of freedom, which is denoted as $\chi^2(\nu)$. Also used in queueing theory.

Normal:

- $f(x) = \dfrac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/\left(2\sigma^2\right)}$, $-\infty < x < \infty$.
- Mean: $\mu = \mu$
- Variance: $\sigma^2 = \sigma^2$
- Application: Many, e.g., random error. Also, a distinguished distribution by way of the Central Limit Theorem.

Student's t:

- $f(x) = \dfrac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)} \dfrac{1}{\sqrt{\nu\pi}} \left(1 + \dfrac{x^2}{\nu}\right)^{-(\nu+1)/2}$, $-\infty < x < \infty$. Here ν is the number of degrees of freedom.
- Mean: $\mu = 0$
- Variance: $\sigma^2 = \frac{\nu}{\nu-2}$
- Application: Approaches a standard normal distribution as $\nu \to \infty$. Also, given n independent samples of normally distributed r.v.'s with a common unknown standard deviation σ, let the sample mean be given by $\overline{x} = (x_1 + x_2 + \cdots + x_n)/n$ and the

sample variance by $s^2 = \dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n}(x_i - \overline{x})^2$, then the random variable

$$t = \frac{X - \overline{X}}{S/\sqrt{n}}$$

has a Student's t distribution with $n-1$ degrees of freedom.

Lognormal:

- $f(x) = \dfrac{1}{\sqrt{2\pi}\sigma x} e^{-(\ln x - \nu)^2/\left(2\sigma^2\right)}$, $-\infty < x < \infty$.
- Mean: $\mu = e^{\nu + \sigma^2/2}$
- Variance: $\sigma^2 = e^{2\nu+\sigma}\left(e^{\sigma^2} - 1\right)$
- Application: This is really the density function for e^X, where X is normally distributed. Equivalently, $X = \ln Y$. It is fundamentally important in modeling the dynamics of asset prices.

Joint Distributions. For the most part, we'll restrict our discussion of joint random variables to continuous distributions, though all the ideas have a discrete counter part. Likewise, the ideas we are going to discuss extend to any number of r.v.'s, like $X_1, X_2, \ldots, X_n$, but we will mostly confine our attention to two r.v.'s, say X and Y.

In order to motivate the idea of joint distributions, let's consider Example 0.5 with a twist: we will throw a dart at our one dimensional dart board twice. With each throw, we will note the position of the outcome on the interval $[0,1]$ and this number is our random variable. This gives us two random variables X and Y which share the same sample space of outcomes when viewed individually. Moreover, it makes the statistics of the new experiment more complicated than just numbers on the interval $0 \le x \le 1$. Now they are ordered pairs of numbers (x, y) such that $0 \le x, y \le 1$; In other words, they belong to a unit square in the xy-plane. The event $X + Y \le 1$ can now be pictured as a subset of this square.

Now suppose we ask the question: what is the probability that $X + Y \le 1$? In order to answer this question, we need to understand how these variables behave jointly, so we will need a p.d.f. $f(x, y)$ that is a **joint distribution** of both random variables. Here "density" means probability per unit area, not length. Once we have such a function, we can describe the probability of an event A occurring as a double sum in the case of discrete r.v.'s and as a double integral in the case of a continuous r.v. Thus,

$$P(A) = \iint_A f(x, y)\, dA.$$

In most cases we can reduce these double integrals over plane regions to iterated integrals as in ordinary calculus. As an example of this, we can define a **joint cumulative distribution function** (c.d.f.) by the formula

$$F(x, y) = P(X \le x, Y \le y)$$

and obtain that

$$F(x, y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f(x, y)\, dy\, dx.$$

Now what about the p.d.f. of the example we have mentioned. This can get complicated. If both throws are random, and the p.d.f. for each r.v. separately is the uniform distribution, it is reasonable to expect that the joint p.d.f. should also be uniformly distributed, so we have $f(x,y)=1$. But what if the throws are not independent? For example, if we play a game where the "score" of the throws, $x+y$, is close to a certain number, then where the first dart landed will affect where we throw the second one. So in this case we would expect $f(x,y)$ to express a more complicated relationship between x and y.

Standard Notation: To save ourselves the inconvenience of always having to assign a name to the joint p.d.f. and c.d.f. of given r.v.'s X and Y, we adopt the convention that

$$\begin{aligned} f_{X,Y}(x,y) &= \text{joint p.d.f. of the r.v.'s } X,Y. \\ F_{X,Y}(x,y) &= \text{joint c.d.f. of the r.v.'s } X,Y. \end{aligned}$$

Expectation and Covariance. Just as with p.d.f.'s of one variable, one can define some key concepts for r.v.'s X and Y:

- **Expectation** of a function $g(x,y)$ of r.v.'s:

$$E[g(X,Y)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)\, f_{X,Y}(x,y)\, dy\, dx.$$

- **Covariance** of X and Y: This is just

$$\operatorname{Cov}(X,Y) = E[(X-E[X])(Y-E[Y])].$$

One can show

$$\begin{aligned} \operatorname{Cov}(X,X) &= \operatorname{Var}(X) \\ \operatorname{Cov}(X,Y) &= E[XY]-E[X]E[Y] \\ \operatorname{Cov}(X,Y) &= \operatorname{Cov}(Y,X) \\ \operatorname{Var}(X+Y) &= \operatorname{Var}(X)+\operatorname{Var}(Y)+2\operatorname{Cov}(Y,X) \end{aligned}$$

- **Correlation** of X and Y:

$$\rho(X,Y) = \frac{\operatorname{Cov}(X,Y)}{\sqrt{\operatorname{Var}(X)\operatorname{Var}(Y)}}.$$

One can show that $-1 \le \rho(X,Y) \le 1$ and that $\rho(X,X)=1$. If $\rho(X,Y)=0$ or, equivalently, $\operatorname{Cov}(X,Y)=0$, we say that X and Y are **uncorrelated**.

- **Independent** r.v.'s X and Y: means that for all a and b,

$$P(X\le a, Y\le b) = P(X\le a)\,P(Y\le b).$$

One can show that if X and Y are independent and $g(x), h(y)$ are any functions then

$$\begin{aligned} F_{X,Y}(x,y) &= F_X(x)\,F_Y(y) \\ f_{X,Y}(x,y) &= f_X(x)\,f_Y(y) \\ E[g(X)h(Y)] &= E[g(X)]\,E[h(Y)]. \end{aligned}$$

Thus we see that if two random variables are independent, then the sum of their variances behaves nicely. In this case we obtain that

$$\operatorname{Var}(\alpha X+\beta Y) = \alpha^2\operatorname{Var}(X)+\beta^2\operatorname{Var}(Y).$$

Multivariate Normal Distributions. Here is a generic example of an extremely important multivariate distribution. In the case of two r.v.'s this type of distribution is called a **bivariate** distribution. These distributions are the "correct" analog in higher dimensions to the normal distributions in one dimension. In the following example we need the concept of a "**symmetric positive definite** matrix (SPD)". First, a square $n \times n$ matrix A is **symmetric** if $A^T = A$. Secondly, A is **positive definite** if $\mathbf{x}^T A\mathbf{x} > 0$ for all nonzero vectors $\mathbf{x}$. Some useful facts:

- If A is symmetric, then all the eigenvalues of A are real.
- A symmetric matrix is positive definite if and only of all its eigenvalues are positive.
- If A is symmetric, then there exists an orthogonal matrix Q (i.e., $Q^T = Q^{-1}$) such that $Q^T AQ$ is diagonal and moreover the diagonal elements are exactly the eigenvalues of A.

Example 0.6. Suppose that we are given a vector μ and an $n \times n$ matrix $C = [c_{i,j}]$ that is SPD. Define the function

$$f(x_1, x_2, \ldots, x_n) = \frac{1}{(2\pi)^{n/2}\sqrt{\det(C)}} e^{-(\mathbf{x}-\mu)^T C^{-1}(\mathbf{x}-\mu)/2}, \ -\infty < x_i < \infty.$$

Then the following are true:

- The function $f(x_1, x_2, \ldots, x_n)$ is a joint p.d.f. for some r.v.'s $X_1, X_2, \ldots, X_n$.
- Each X_i is normally distributed with mean μ_i and variance $c_{i,i}$.
- $c_{i,j} = \operatorname{Cov}(X_i, X_j)$.

The following theorems can be established with some work:

Theorem: If $\mathbf{X}$ is a multivariate normal random vector with expected values given by the vector μ and covariance matrix C , and if $\mathbf{Y} = A\mathbf{X}$ then $\mathbf{Y}$ is also multivariate with $E[Y] = A\mu$ and $\operatorname{Cov}(\mathbf{Y}) = ACA^T$.

Theorem: If $\mathbf{X}$ is a multivariate normal random n-vector with expected values given by the vector μ and covariance matrix C of full rank, then the r.v.

$$Z = (\mathbf{X} - \mu)^T C^{-1} (\mathbf{X} - \mu)$$

has a chi-square distribution with n degrees of freedom.

Parameter Estimation

Point Estimation. Here is the problem: We are given independent r.v.'s $X_1, X_2, \ldots, X_m$ with p.d.f.'s $f_1(x_1), f_2(x_2), \ldots, f_m(x_m)$. However, the p.d.f.'s are complicated by the fact that they in turn depend on n parameters $m_1, m_2, \ldots, m_n$ which can be viewed as the components of the parameter vector $\mathbf{m}$. The problem is that we don't know the parameters. How do we take the outcome of an experiment, say $X_1 = x_1, X_2 = x_2, \ldots, X_m = x_m$ and use the joint p.d.f. of this distribution to say something about $\mathbf{m}$? Specifically, we would like a **point estimate** of $\mathbf{m}$.

One approach is to use a **maximum likelihood estimator (m.l.e.)** of $\mathbf{m}$. Here's how it is done:

(1) Form the joint p.d.f. of these independent r.v.'s, which we know is,

$$f(x_1, x_2, \ldots, x_m|\mathbf{m}) = f_1(x_1|\mathbf{m}) f_2(x_2|\mathbf{m}) \cdots f_m(x_m|\mathbf{m}).$$

Here the notation acknowledges that f depends on a given $\mathbf{m}$.

(2) View f as a function of $\mathbf{m}$ and rename it as the **liklihood function**

$$L(\mathbf{m}) = f(x_1, x_2, \ldots, x_m|\mathbf{m})$$

(3) Find the value of $\mathbf{m}$ that maximizes the function for the particular sample $x_1, x_2, \ldots, x_m$. This is the maximum likelihood estimator of the parameter in question.

Note: this usually involves setting the gradient of L equal to zero and solving the resulting system. This system is guaranteed to have a maximum if, e.g., f is continuous. For f is non-negative everywhere and tends to zero as $\|\mathbf{m}\| \to \infty$.

Example 0.7. Each X_i has distribution $N(d_i, \sigma_i)$, i.e., is normally distributed with mean μ_i and variance σ_i. Suppose further that

$$\mu = [\mu_1, \mu_2, \ldots, \mu_m]^T = G\mathbf{m}$$

Find a maximum likelihood estimator for $\mathbf{m}$.

Solution. Form the joint pdf and obtain

$$L(\mathbf{m}) = \frac{1}{\sqrt{2\pi}\sigma_1} e^{-(x_1-\mu_1)^2/\left(2\sigma_1^2\right)} \cdots \frac{1}{\sqrt{2\pi}\sigma_m} e^{-(x_m-\mu_m)^2/\left(2\sigma_m^2\right)}.$$

In this case we will not follow the procedure above because it is about as much work as we had to do in deriving the equivalence of the normal equations and the least squares problem. Let's just take a close look at what we're maximizing and obtain that a m.l.e. of $\mathbf{m}$ comes from a least squares solution to the system.

$$WG\mathbf{m} = W\mathbf{d}$$

where $W = \operatorname{diag}(1/\sigma_1, \ldots, 1/\sigma_m)$.

Here are two desirable properties of an estimator:

(1) Consistent (converges in probability)
(2) Unbiased estimator.

Confidence Intervals. The word "statistic" has different meanings to different folks. For us a **statistic** shall mean any definite function of one or more r.v.'s. Two important examples come from the notion of a random sample, which means a sequence of independent and identically distributed (abbreviated to i.i.d.) random variables:

- The **mean** of a random sample $X_1, X_2, \ldots, X_n$:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

- The **variance** of a random sample $X_1, X_2, \ldots, X_n$:

$$V^2 = \frac{1}{n}\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2.$$

- The **sample variance** of a random sample $X_1, X_2, \ldots, X_n$:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2.$$

Key question: How do we estimate the mean and and variance of the distribution of these random samples, or for that matter any other parameter associated with the distribution?

This question leads us to the notion of **confidence intervals**: given a probability $1-\alpha$ called a "**confidence coefficient**" use the data to construct the smallest possible interval I of real numbers such that the probability that the true value of the parameter being in this interval is $1-\alpha$. The use of $1-\alpha$ is a matter of convenience in formulas. One ofter sees terms like the "95% confidence interval". This means the confidence interval found with $1-\alpha=0.95$. We'll explore these ideas for the case of mean and variance.

Motivation: What's so great about the mean?

Example 0.8. Suppose that we are attempting to measure a ideal and definite physical quantity, say the mass m of the object. We do so by taking repeated measurements of the mass, say $m_1, m_2, \ldots, m_n$. What to do with these numbers? In the absence of any other information, we might average them out, in the hopes that errors will somehow cancel each other out. Is this realistic? Answer: sometimes.

Specifically, we'll describe the experiment more formally as a sequence $M_1, M_2, \ldots, M_n$ of r.v.'s. We may write

$$M_i = m + X_i$$

where X_i is the error of the ith measurement. Certainly, it is reasonable to assume that these r.v.'s are independent. In many cases it is also reasonable to assume that the errors are normally distributed with mean 0 and standard deviation σ. It follows that the M_i are normally distributed with mean

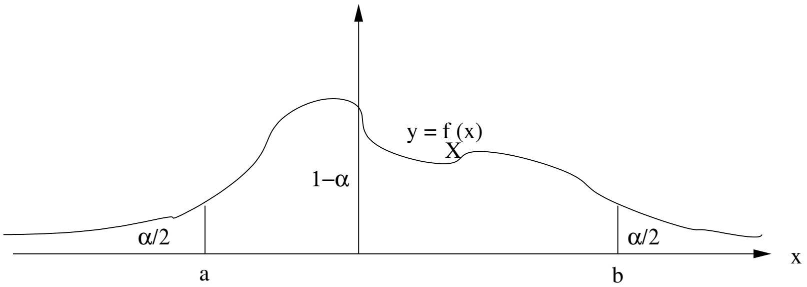
$$E[M_i] = E[m + X_i] = E[m] + E[X_i] = m + 0 = m$$

and variance

$$\mathrm{Var}(M_i) = \mathrm{Var}(E_i) = \sigma^2.$$

In particular, the number we're really interested in, m, is the expectation of a normal random variable. We'll see why the sample mean is useful by answering the key question.

The basic idea for computing confidence intervals is to find a statistic X that is a "good" estimator of the desired parameter and has a known distribution, which can be used to compute a confidence interval. Here's how: we split the probability into half and construct an interval which has $\alpha/2$ area under the p.d.f. f_X to the left of a point a and area $\alpha/2$ under the p.d.f. to the right of a point b. Here's a picture:



We accomplish locating the points a and b as follows: solve the equations

$$\begin{aligned} F_X(a) &= \frac{\alpha}{2} \\ F_X(b) &= 1 - \frac{\alpha}{2}. \end{aligned}$$

If F_X is continuous, we are guaranteed that solutions exist, and in fact, an inverse function to the c.d.f. F_X exists.

Finally, WHAT DOES ALL THIS MEAN, exactly??? It means that if you calculate a confidence interval based on data you have observed, and if all the hypotheses about i.i.d. normal r.v.'s is correct, then the true value of parameter you are estimating is in this interval with a probability of $1 - \alpha$. Put another way: $100 \cdot (1 - \alpha)$ times out of 100 this calculation will yield an interval containing the desired parameter.

.

Estimating Mean with Known Variance and Normal Distribution. Some simple facts about normal distributions play a key role here. Suppose that $X_1, X_2, \ldots, X_n$ are i.i.d. normal r.v.'s with mean μ and variance σ^2. From various facts outlined in these notes we have:

- $X_1 + X_2 + \cdots + X_n$ has a normal distribution with mean $n\mu$ and variance $n\sigma^2$. (See p. .)
- So $\frac{1}{n}(X_1 + X_2 + \cdots + X_n) = \overline{X}$ has a normal distribution with mean $n\mu/n = \mu$ and variance $n\sigma^2/n^2 = \sigma^2/n{=}(\sigma/\sqrt{n})^2$. (See p. .)
- Hence $Z = \left(\overline{X} - \mu\right) / (\sigma/\sqrt{n})$ has a standard normal distribution. (See p .)

Thus we have shown that

Theorem. *Let $X_1, X_2, \ldots, X_n$ be i.i.d. normal r.v.'s with mean μ and variance σ^2. Then the statistic*

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

has a standard normal distribution.

Estimating Mean with Unknown Variance and Normal Distribution. The Student's t distribution plays a key role here. The key theorem is as follows:

Theorem. *(Sampling Theorem) Let $X_1, X_2, \ldots, X_n$ be i.i.d. normal r.v.'s with mean μ and variance σ^2. Then the statistic*

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

has a Student's t distribution with $n - 1$ degrees of freedom.

Estimating Variance with a Normal Distribution. The chi-square distribution plays a key role here.

Theorem. *Let $X_1, X_2, \ldots, X_n$ be i.i.d. normal r.v.'s with mean μ and variance σ^2. Then the statistic*

$$Y = (n-1)\frac{S^2}{\sigma^2}$$

has a chi-square distribution with $n - 1$ degrees of freedom.

We use these statistical facts as outlined above. If we are given sample data, we calculate the resulting test statistic and observe whether or not it falls in the confidence interval.

Test of Hypotheses. Confidence intervals give us one tool for making decisions. They can be placed in the broader context of *testing of hypotheses.* For example, a manufacturer might claim that a beam has load capacity of $\mu_0 = 200$ tons. We don't really care if the capacity is really greater than 200, but we certainly do care if it is less.

Here is an outline of how a typical testing of hypothesis unfolds in general:

(1) Formulate formulate two hypotheses, the so-called *null hypothesis* H_0, and an alternative hypothesis H_1. Here the hypotheses are exclusive, but not necessarily exhaustive.

(2) Now we pick a significance level α for the test (which is the complement of the confidence level α we used in the previous section.) Typical choices are $\alpha = 0.05, 0.01, 0.001$, or, if you prefer percentages, 5%, 1% or 0.1%. Once we have chosen α, we can compute the **critical region**, that is, the range of values of the sample statistic for which we will reject the null hypothesis. We choose the region so that the probability of the rejecting the null hypothesis when it is true is exactly α. The complement of the critical region is the **acceptance region**.

(3) Choose a random variable with known distribution which depends on the hypotheses,

$$\widehat{\Theta} = g\left(X_1, X_2, \ldots, X_n\right)$$

and that gives a "good" estimator for where the observed parameter θ should be. We use this statistic to determine a critical region.

(4) Use a sample $x_1, x_2, \ldots, x_n$ to determined an observed value $\widehat{\theta} = g\left(x_1, x_2, \ldots, x_n\right)$.

(5) Accept or reject the null hypothesis based on whether or not the observed value $\widehat{\theta}$ falls in the critical region.

Notice that there are two types of error that we could commit:

(1) **Type I error:** reject H_0 when it is true. In this case we know exactly what the likelihood of error is, namely, it is just the significance level α of our test.

(2) **Type II error:** accept H_1 when it is false. We label the probability of this type of error as β. The number $1 - \beta = \eta$ is called the **power** of the test. This number isn't obvious as α as that it actually depends on the true value θ of the parameter as well as the hypothesized value θ_0. In fact, we can write $\eta = \eta\left(\theta\right)$ and $\beta = \beta\left(\theta\right)$.

(3) Types I and II errors are related. Generally, for a fixed sample size reducing the probability of one will increase the other. Given specific values of the true θ and a fixed α, we can estimate the sample size needed to achieve a given power η. Notice that if the true value θ is close, but not equal to the hypothesized value θ_0, then β approaches α, so the power of the test approaches zero. In general, the only way to reduce both is increase the sample size.

Consider the example we started above:

Example 0.9. We want to test a manufacturer's claim that a certain type of beam that they produce has load capacity of $\mu_0 = 200$ tons. If the capacity is significantly less that 200, we will not purchase the beams. We let μ be the true value of the parameter. Hence, we formulate these hypotheses:
$H_0 : \mu = \mu_0 \equiv 200$.
H_1: $\mu = \mu_1 < 200$.
We'll assume that the individual capacities are normally distributed about the mean μ. Next,

since we don't know the variance of this distribution, we select as an estimator statistic

$$\widehat{\Theta} = T = \frac{\overline{X} - \mu}{S/\sqrt{n}} = T(\mu).$$

We know that T has a Student's t distribution with $n-1$ degrees of freedom by our assumption on the individual capacities. Finally, we compute the sample statistic under the assumption $\mu = 200$ to obtain a value $\widehat{\theta}$. If this value falls in the critical region we reject the null hypothesis. Otherwise we accept it. Actually, a better phrase than "accept it" might be that we do not reject the null hypothesis. In this situation it's easy to calculate the smallest value of the power function, namely

$$P\left(\widehat{\Theta} \leq c\right) = F_{T(\mu)}(c)$$

where

$$1 - \alpha = F_{T(\mu_0)}(c).$$