# Highlights of Basic Probability and Mathematical Statistics

## Executive Summary

This document is a summary of some important basic facts from probability and mathematical statistics, organized into six sections:

1. Probability
2. One random variable
3. Multiple random variables
4. Random sample model
5. Linear combinations
6. Estimation

Key concepts and points:

- Probability measures the degree of (un)certainty.
- There are two schools of thought on the interpretation of probability: objective (frequentist) and subjective (Bayesian).
- A random variable is a set of potential outcomes and a function to calculate their probabilities.
- Random variable *X* vs. realization *x*: Random variable *X* is potential value; realization *x* is actual value.
- Random Sample model (RS) is a particular set of specifications for a sample: Each selection comes from the same population and the selections are mutually independent – independent and identically distributed (i.i.d.).
- Mean and variance of a linear combination, covariance between two linear combinations.
- The ordinary least squares (OLS) estimate is the parameter value that is closest to the data collectively. I.e., the OLS estimate is the value of the model specification $g(\theta_1,\theta_2,...,\theta_p)$ that minimizes $\sum_{i=1}^{n}\{x_i - g(\theta_1,\theta_2,...,\theta_p)\}^2$.
- The maximum likelihood estimate (MLE) is the set of parameter values that makes the observed data most likely. I.e., the MLE is the value of $\theta_1,\theta_2,...,\theta_p$ that maximizes the likelihood (the density as a function of the parameters), given data realizations $x_1,x_2,...,x_n$:

$$\sup_{\theta_1,\theta_2,...,\theta_p} L(\theta_1,\theta_2,...,\theta_p;x_1,x_2,...,x_n) = \sup_{\theta_1,\theta_2,...,\theta_p} f(x_1,x_2,...,x_n;\theta_1,\theta_2,...,\theta_p).$$

- The Central Limit Theorem (CLT) for Random Samples: If $X_1, X_2,...,X_n$ is a RS and the common mean and standard deviation are $\mu$ and $\sigma^2$, respectively, then $P\left(a < \dfrac{\overline{X}-\mu}{\sigma/\sqrt{n}} < b\right)$ gets close to $P(a < N(0,1) < b)$ as *n* gets big, for all fixed *a* and *b*.
- How Chi-square, T, and F random variables are obtained from normal random variables.
- A confidence interval is a random interval $(g_1(X_1,...,X_n), g_2(X_1,...,X_n))$ with statistics as endpoints such that there is a calculable probability (usually high) that the interval contains a parameter: $P(g_1(X_1,...,X_n) < \theta < g_2(X_1,...,X_n)) = 0.95$ (for example).
- An hypothesis is a statement about a parameter that is either true or false. An hypothesis test is a set of outcomes *R* defined by a statistic $T(X_1,...,X_n)$ such that the decision to

reject the hypothesis is taken if $T(X_1,...,X_n) \in R$. For example, reject $H_0 : \mu \leq 6$ if $\bar{X} > 8$.

------------------------------------------------------------

# Highlights of Basic Probability and Mathematical Statistics

## Section 1. Probability

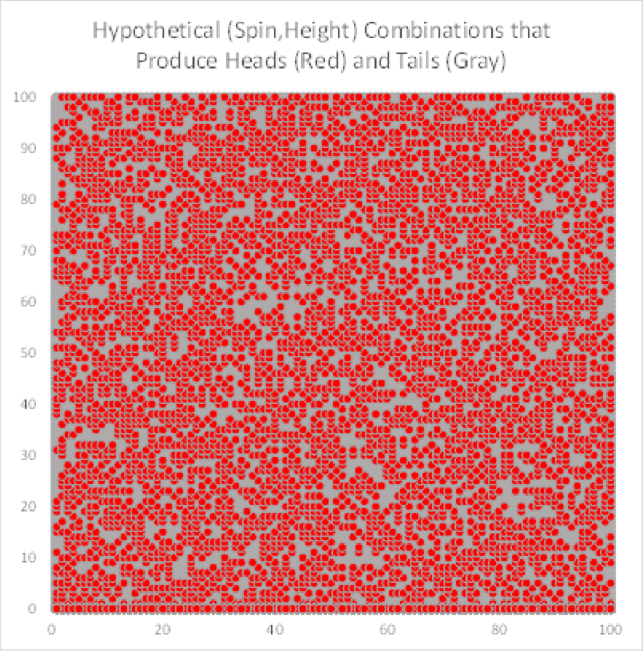
### Two Views of the Nature of Probability

Fundamentally, statistics is about managing uncertainty, and probability is a fundamental measure of uncertainty. There are two major philosophical interpretations of uncertainty, hence of probability, hence of statistics. These views may be labeled the *objective* and *subjective* views – alternatively, often labeled the *frequentist* and *Bayesian* views, respectively. It is important to understand the differences between them in broad outline, because they lead to different statistical practice in managing uncertainty. However, they use the same foundational mathematics. I will present an outline of the foundational mathematics of probability. But first, a short overview of these two views of probability.

I will illustrate the two views with a canonical example of a random event – the tossing of a coin. When we toss a coin, there are two possible outcomes: head (1) and tail (0). Before we toss the coin, we are uncertain which outcome will occur. We say that the outcome is "random" or determined by "chance". Actually, it is not. The outcome is determined by the laws of physics in conjunction with the amount of spin that we impart to the coin, the height we toss it, the level at which we catch it, whether we turn it over or leave it up in our hand – conceivably even the air pressure and humidity. If we knew physics sufficiently well and could measure the relevant input factors sufficiently well, then we could – in principle – calculate the outcome as soon as the coin is released into the air. If the coin outcome is determined, what then do we mean by the *probability* that the outcome will be a head? Although the outcome may be physically determined, we are still uncertain because we do not know physics well enough and cannot measure the relevant factors well enough. The two views of probability that I mentioned *interpret* the meaning of that uncertainty for us.

**The objective view.** When we toss a coin, we select a combination of spin, height, etc. that leads to a head – select a slightly different combination and we get a tail. For a fair coin, about half of those possible combinations lead to heads and about half lead to tails. The "head" input factor combinations are so thoroughly mixed with the "tail" input factor combinations that we cannot control which group we pick from. It is futile to try. Toss the coin again and we blindly choose either one of the combinations that become a head or one of the combinations that become a tail. This situation is illustrated in Figure 1 (below). The red dots represent combinations of spin and height that, if selected, will produce a head; the gray dots represent combinations that will produce a tail. About half the dots are red and about half are gray, and they are so thoroughly mixed that it is futile to try to shoot for a particular color. Now toss repeatedly. Since head and tail combination types are equally numerous for a fair coin and well-mixed, about half of our choices will be from the "head" combinations and about half from the "tail" combinations. *Probability* refers to the number of "head" combinations in relation to the number of "tail" combinations. If we change the physical characteristics of the coin, say by plating the tail side with lead, then many combinations that formerly would have become tails now become heads (because the coin is now bottom-heavy and more likely to land face-side up). The probability of heads increases. Objectively, we can discover the probability of heads to any desired degree of accuracy by repeatedly tossing the coin and observing the relative frequency of heads. Equivalently, we are actually discovering empirically the relative frequencies of the red and gray dots. Hence, the objective view is also labeled the frequentist view.

Figure 1



**The subjective view.** The real world *knows* what it needs to do. That is given by physical law. Uncertainty is a mental state. Therefore, probability is also a mental state. When a coin is tossed in *exactly* the same way, exactly the same outcome will occur. Variation in outcome occurs only if the input factors are varied. What we discover by repeatedly tossing the coin is how one person varies the input factors. Another person may vary them in a different manner – e.g., slow spin versus fast spin. Who is to say that the input conditions leading to heads are equally mixed under all tossing conditions for all people? That is an opinion, or an assumption, which is a mental state. What matters is not whether there is some objective probability that we can discover after sufficient experimentation. What is important is how we should change our uncertainty in response to additional data, however many or few there may be. For example, we may believe initially that the coin is fair. However, if we start tossing the coin and we observe that the outcome heads builds up a consistent lead over tails, then we may change our minds. The subjective approach provides a mathematically based method (Bayes theorem) for rationally modifying our probability assessments as new data become available. One person may think that heads is more likely than tails, and another person may think that tails is more likely than heads. Each would have a different probability for heads – and that is OK, as long as each behaves rationally (per Bayes theorem) to revise his/her probabilities as new data come in from repeated coin tosses.

The objective and subjective views differ substantially in their attitudes toward parameters. These attitudes lead to differences in statistical practice. A parameter is an unknown value in a statistical model. For example, in the case of tossing a coin, the key parameter is $\pi$ = probability of heads. The value of $\pi$ must lie between 0 and 1. Both frequentists (objectivists) and Bayesians (subjectivists) use parameters like $\pi$. However, to a frequentist, $\pi$ is a fixed, definite number – a constant – but we do not know what its value is. If the coin is fair, then $\pi$ = 0.5. However, the coin may be imperfect. So we may toss the coin repeatedly to learn, to desired levels of precision, what the value of $\pi$ really is. To a Bayesian, $\pi$ is a random variable. This means that the Bayesian assesses her uncertainty about $\pi$ in the form of a probability distribution with a probability density function $f(\pi)$ and cumulative distribution function $F(\pi)$. For example, a Bayesian may assess a small probability (say, 0.01) that the probability of heads is below 0.4 [F(0.4) = 0.01], a large probability that $\pi$ is between 0.4 and 0.6 [say, F(0.6) – F(0.4) = 0.98], and a small probability that it is above 0.6 [1 – F(0.6) = 0.01]. To say that $\pi$ is a random variable does not necessarily mean that the value of $\pi$ actually *varies* from toss to toss (although some Bayesians are philosophically indifferent whether or not $\pi$ varies). It just means that the Bayesian's mental state of uncertainty about $\pi$ can be expressed by a distribution $f(\pi)$. The Bayesian will modify $f(\pi)$ in accordance with Bayes theorem as data from repeated coin tosses accumulate. For the frequentist, statistical inference is about sharpening objective knowledge about what the value of $\pi$ really is. For the Bayesian, statistical inference is about refining her mental state of uncertainty about $\pi$.

Fortunately, whether a statistician subscribes to the objective or to the subjective interpretation of probability, the mathematical theory and laws of probability are the same.

## Foundations of Probability

Probability begins with a set U of all possible outcomes of an observation or experiment. U is called the **universe** or **population**. An example is U = {1,2,3,4,5,6}, which is all of the possible outcomes of throwing a die. A probability measure is a function *P* that assigns numbers, called **probabilities**, to subsets of U. The subsets are called **events**. The probabilities must lie between 0 and 1 inclusively. Moreover, the probabilities must satisfy certain consistency conditions. The following is a list of the conditions that a probability function must satisfy:

1. P(U) = 1
2. P(A$^c$) = 1 – P(A) where A$^c$ = U – A is the complement of A in U.
3. P($\bigcup_{n=1}^{\infty} A_n$) = $\sum_{n=1}^{\infty} P(A_n)$ for any collection of disjoint subsets $A_1, A_2, \ldots$ of U.

For example, suppose that a fair die is thrown. Then each of the six possible outcomes is assigned probability 1/6. Then

1. P("throw 1,2,3,4,5, or 6") = P({1,2,3,4,5,6}) = 1
2. P("throw even") = 1 – P({1,3,5}) = 1 - 3/6 = 3/6
3. P("throw even or 3") = P({2,4,6}) + P({3}) = 3/6 + 1/6 = 4/6

From the consistency conditions, other rules for working with probabilities can be derived. For example:

Union. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

If A and B are disjoint then $P(A \cup B) = P(A) + P(B)$

Intersection. $P(A \cap B) = P(A) + P(B) - P(A \cup B)$

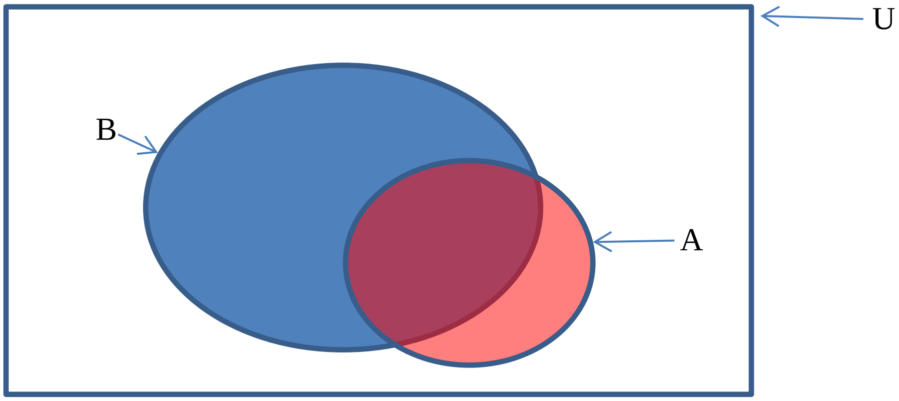
If A and B are disjoint then $P(A \cap B) = 0$

The concept of conditional probability plays an important role in both frequentist and Bayesian statistics. Here is the definition:

Definition. The **conditional probability of A given B** is P(A | B) = $\frac{P(A \cap B)}{P(B)}$, if P(B) ≠ 0.

What this means is that if B is given – that is, if B is known to occur – then the rest of U is no longer relevant for the calculation of the probability of A. U – B is discarded, and B becomes the new universe. The probability of A becomes the part of A that remains, relative to B. The following Venn diagram illustrates the situation.

Figure 2



Suppose that area represents probability. Then the area of the rectangle U is 1. The probability of A is the area of the entire oval of A within U. This is the **unconditional probability** of A. If we are given that B occurs, then the part of A that is outside of B – namely, the orange part of A – cannot occur. The only part of A that can occur is the part within B – the overlap, or intersection of A with B. The size of the overlap *relative to B* is the new **conditional probability** of A.

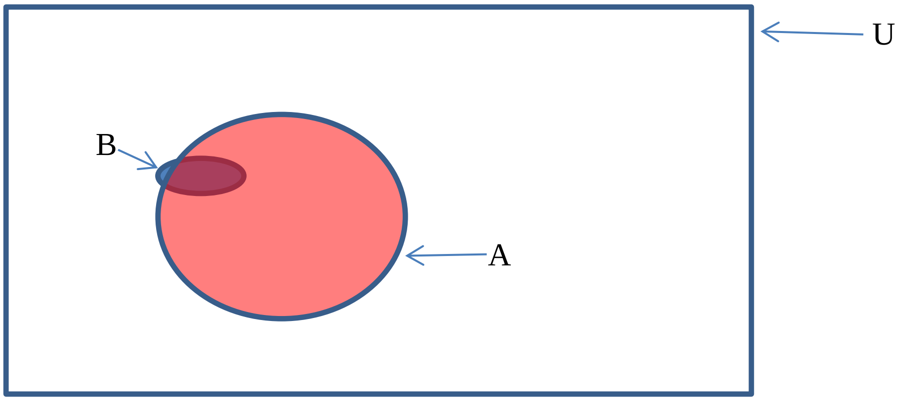
> For example, suppose we throw a fair die. Suppose that B is {2,4,6} and A is {5,6}. Then the unconditional probability of A is 2/6. The unconditional probability of B is 3/6. If B is given to occur, then the 5 in A cannot occur. The new universe becomes B = {2,4,6}. The only part of A left in the new universe is 6. 1/3 of the new universe is A, so 1/3 is the conditional probability of A given B.
>
> The definition of conditional probability allows us to calculate the new probability by using old probabilities in the old universe U: P(A | B) = P({5,6} | {2,4,6}) = P({6}) / P({2,4,6}) = 1/6 ÷ 3/6 = 1/3.

It is very important to note that P(A | B) is not the same as P(B | A). Interchanging these two conditional probabilities is the source of much confusion in thinking about probabilities. Mathematically, the difference is clear.

> For example, in the immediately preceding illustration with outcomes of a fair die roll, P(A | B) = 1/3, but P(B | A) = ½. (Verify this.)

However, in the discourse of ordinary life, the difference is often not clear to a lot of people. For example, if you say "most criminals are poor people," some become incensed because they think that you have impugned the moral fiber of the poor. However, you have made a true statement about P(A | B), where A = poor, B = criminals. You have correctly said that the probability of being poor, given that a person is a criminal is high. You have not made a statement about P(B | A) = probability of being a criminal given that a person is poor. The former is large, the latter small. This situation is shown in the following Venn diagram.

Figure 3



What if it turns out that P(A) = P(A | B)? Then the probability of A is unaffected by whether we know that B has occurred or not. The probability of A remains the same. The probability of A is independent of B.

Definition. Events A and B are **independent** if P(A | B) = P(A). Otherwise, A and B are **dependent**.

Note: If P(A | B) = P(A) then it is also true that P(B | A) = P(B), and the complements $A^c$ and $B^c$ are also independent. So there is no ambiguity in saying "A and B are independent." A is independent of B if and only if B is independent of A, and likewise for the complements.

The definition of conditional probability yields another rule for calculating probabilities of intersections:

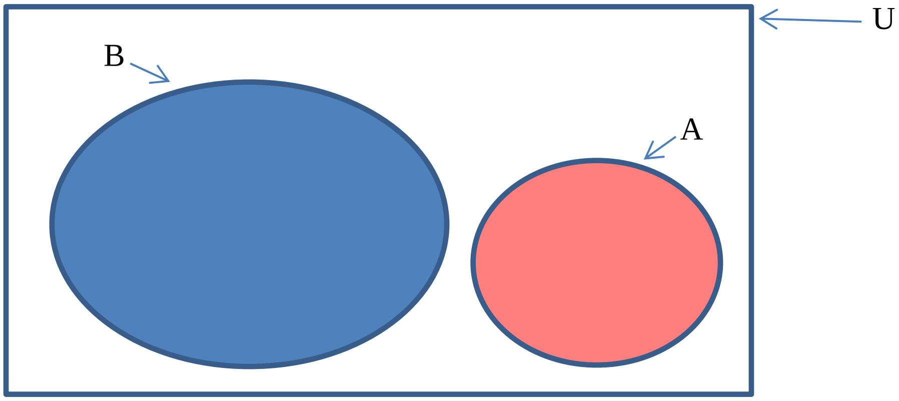
Intersection. $P(A \cap B) = P(A)P(B \mid A) = P(B)P(A \mid B)$

If A and B are independent, then $P(A \cap B) = P(A)P(B)$

The notion of independent events can be illustrated by a Venn diagram. Consider Figure 2 above. If the area of A in B is the same as the area of A in U, then the probability of A remains the same when the universe is shrunk from U to B. This is required for independence. In Figure 3, when the universe is shrunk from U (everyone) to B (criminals), the relative size of A goes from small to huge. Criminals and poor are dependent events.

Students sometimes think that disjoint events are independent. After all, disjoint events have nothing in common. Figure 4 illustrates the situation. In Figure 4, A has positive area in the universe U. But A has no overlap with B. So if we know that B occurs, then A cannot occur. A goes from positive area in U to zero area in B – from positive probability in U to zero probability in B. In the example of the roll of a fair die, suppose A = {2,4,6} and B = {1,3,5}. Then P(A) = 3/6. But P(A | B) = 0. In order for events A and B to be independent, A must maintain the same relative representation in B as it has in the universe U.

Figure 4



**Bayes Theorem.** $P(B\,|\,A) = \dfrac{P(A\,|\,B)P(B)}{P(A\,|\,B)P(B) + P(A\,|\,B^c)P(B^c)}$

Bayes Theorem provides a way to calculate the *reverse* conditional probability P(B | A), if you are given P(A | B) and related probabilities. The formula looks formidable. But it can be deconstructed into parts that make intuitive sense. From the definition of conditional probability, the numerator is just $P(A \cap B)$. The denominator is just P(A), written as two parts: the part of A in B [i.e., $P(A \cap B)$ ] and the part of A not in B [i.e., $P(A \cap B^c)$ ]. The power of Bayes Theorem will be more apparent after considering a real example:

Ex: The June, 2002, issue of Scientific American (p.16) reported a new test for ovarian cancer, which strikes about 1 in 2500 women over the age of 35 in the U.S. each year. In developing the test, researchers gave the test to a group of women known to have ovarian cancer. The test was positive on all of these women. The researchers also gave the test to a group of women known not to have ovarian cancer. The test was negative on 95% of them. The developmental test results sound great. But should the test be used to screen women for ovarian cancer? The article did not address this question, nor did it do the calculation that I provide below. But the information in the article can be formulated in a way that Bayes Theorem can provide a definitive and surprising answer.

Let B be the event that a woman over the age of 35 in the U.S. has ovarian cancer. Let A be the event that a woman over the age of 35 in the U.S. has a positive result on the test (the test says cancer). Then from the information in the article, we have

$$P(B) = 1/2500 = 0.000400; \quad P(A\,|\,B) = 1.000; \quad P(A^c\,|\,B^c) = 0.95$$

And from the latter, we have $P(A\,|\,B^c) = 0.05$. Reasoning with conditional probabilities can be tricky. So if these probabilities are not clear to you, please review them carefully.

What doctors really want to know is P(B | A) = the probability that a woman actually has ovarian cancer if the test predicts that she does. That figure was not given in the article but can be calculated from Bayes Theorem:

$$P(B\,|\,A) = \frac{P(A\,|\,B)P(B)}{P(A\,|\,B)P(B) + P(A\,|\,B^c)P(B^c)} = \frac{1.000 * 0.000400}{1.000 * 0.000400 + 0.05 * 0.999600} = 0.007940$$

Fewer than 1 in 100 women who test positive will actually have ovarian cancer. The test should not be used to screen women for ovarian cancer.

What happened? All of the developmental statistics sounded wonderful. Insight can be gained by looking at the expected counts in a classification table. Suppose 1,000,000 women over 35 are given the test. From the figures given in the article, about 1/2500 * 1,000,000 = 400 will get ovarian cancer. Of those, all 400 will test positive. 999,600 women will not get ovarian cancer. Of those, 95% will test negative: 0.95 * 999,600 = 949,620; and 0.05 * 999,600 = 49,980 will test positive. We put these counts into a **classification table**:

| | | Truth | | |
|---|---|---|---|---|
| | | Cancer | No Cancer | TOTAL |
| Test | Positive | 400 | 49,980 | 50,380 |
| | Negative | 0 | 949,620 | 949,620 |
| | TOTAL | 400 | 999,600 | 1,000,000 |

Now we see clearly what the issue is. Out of a total of 50,380 women who test positive, only 400 actually have cancer. This is a fraction equal to 400 / 50,380 = 0.007940 – the probability given by Bayes Theorem. The issue is the large number of **false positives**, 49,980, in relation to the **true positives**, 400. There are no **false negatives** – all negative results are **true negatives**, 949,620. This means that if a woman tests negative, then she is certain to not have ovarian cancer. The **sensitivity** is perfect: P(A | B) = 1.000. Sensitivity is the proportion of cases correctly identified. Although the **specificity** is high, i.e. $P(A^c \mid B^c) = 0.95$, B is so rare that even a high specificity results in misclassifying a large number of truly healthy women. Specificity is the proportion of non-cases correctly identified.

The use of classification tables and the related concepts of the preceding paragraph are extremely important in statistical classification methodologies like logistic regression, discriminant analysis, probit analysis and others. Classification is a frequent objective of Big Data and data mining. Classification attempts to predict who will buy your products, whether a borrower will become insolvent, if a new drug works, whether a defendant is guilty. Classification uses a rule based on the characteristics of the individual or thing being classified to predict whether the individual has a trait or not. The efficacy of the classification rule can be evaluated in a number of different ways in which the four key conditional probabilities weigh in. In general, suppose that A means that an individual is classified as having the trait; B means that the individual actually has the trait.

- **P(A | B)** is the **sensitivity** of the rule = the conditional probability that an individual with the trait is correctly classified as having the trait, i.e., how sensitive the rule is to the presence of the trait.
- **$P(A^c \mid B^c)$** is the **specificity** of the rule = the conditional probability that an individual without the trait is correctly classified as not having the trait, i.e., how specifically the rule is tied to the presence of the trait. Specificity is a necessary antidote to overly generous rules. For example, it is always possible to get a rule that scores 100% on sensitivity – just declare everyone to have the trait! Then every B will be an A, so the sensitivity P(A | B) = 1. Such a rule is called a **no-brainer rule** because it takes no brains to come up with it. However, the no-brainer rule is not specific to B, since every non-B will also be classified A. So the specificity $P(A^c \mid B^c) = 0$. A useful rule will have to balance sensitivity against specificity. But good sensitivity and specificity are not enough to have a good classification rule, as we saw in the preceding ovarian cancer example. There are two more considerations:
- **$P(B^c \mid A)$** is the **false positive rate** = the conditional probability that an individual classified to have the trait in fact does not have it. These are mistakes. Sometimes called

the **false alarm rate**, because you can imagine the rule going off like a fire alarm whenever it finds someone that it thinks has the trait. However, there is really no fire. In many cases, the false positive rate is the most important of all these probabilities. It reflects your actual use of the rule: You will look at the prediction made by the rule and what you care about is whether the prediction is correct. Sensitivity and specificity are concerned with pre-use testing of the rule. To test the rule, you will try it out on those known to have the trait and those known not to have the trait and see what predictions the rule makes. But you will actually use the rule in the reverse sense – you will see the prediction and wonder what is the reality. The ovarian cancer example shows that good test results do not guarantee good performance in the field.

- **P(B | $A^c$)** is the **false negative rate** = the conditional probability that an individual classified not to have the trait in fact does have it. These also are mistakes. The false negative rate is a necessary antidote to overly stringent rules. For example, it is always possible to get a rule that scores 0% on the false alarm rate – just declare that everyone does not have the trait! Then there will be no A, so P($B^c$ | A) = 0. Everyone will be an $A^c$, including all B's, so P(B | $A^c$) = 1. Such a rule is also a no-brainer. To get a useful rule, the false positive and false negative rates must be balanced.

The ovarian cancer example also illustrates how a Bayesian would use the information that a woman tests positive: Prior to the test, a woman has a 1/2500 = 0.000400 chance of having ovarian cancer. She is a Bayesian and has reason to believe that her probability of having ovarian cancer is no different from that of any other U.S. woman, aged 35 or more. Her probability is 0.000400. She takes the test and the test is positive. What should she do? As a good Bayesian, she will revise her probability. How? Answer: By Bayes Theorem. After a positive test result, her probability of having ovarian cancer increases nearly 20-fold, from 0.000400 to 0.007940. However, the probability is still very small. If she takes follow-up tests, she will revise her probability further, based on the results of the additional tests. Similarly, if the woman takes the test and the test is negative, she will also change her probability. After a negative test result, her probability of having ovarian cancer drops from 0.000400 to 0. So she is then certain not to have ovarian cancer.

The frequentist statistician agrees with this computation. Bayes Theorem is a true theorem of mathematics. No one can logically disagree with the result. If the Bayesian believes that her (pre-test) **prior probability** of cancer is more or less than the population figure of 1/2500, she is free to use a different figure. The numbers in the classification table would change accordingly, as would the Bayesian update probability of cancer given a positive test.

# Section 2. One Random Variable

A random variable is a model for a process of **observation** or of **experimentation**. One random variable produces one observation or experimental value. It is important to distinguish observation from experimentation. Experimentation puts more under your control than observation does.

> Ex: To find the relationship between area (X) of an apartment and its rent (Y), you probably take a sample of apartments and *observe* both variables. Both are uncontrolled. Both are random. If you take a sample of apartments drawn from those with 400, 600 and 800 square feet, you would be doing an experiment. You *control* area (X) and *observe* rent (Y).
>
> Ex: To find the relationship between dose of a statin drug (X) and blood cholesterol (Y), you probably take a sample of patients and determine the dose level of X to administer to the patient. You *control* the dose X. You then *observe* the cholesterol Y. Only Y is random. You are doing an *experiment*. You probably do not administer a random dose to the patient.

There are important theoretical and practical differences between observational studies and experiments. But both can be modeled by random variables.[1]

The set of all possible outcomes for the observation or experiment is the **population**.

A collection of observations or experimental outcomes is a **sample**. A sample is a subset of the population.

A **random variable** *X* is a population of outcomes {*x*'s} and a probability function $f(x)$ that assigns probabilities to those outcomes. Note that a random variable is defined to be two things together. It is *both* a set of outcomes *and* their probabilities. It is not just the outcomes and not just the probabilities, but both together. The probability function assigns probabilities in one of two ways, depending on whether *X* is discrete or continuous (see below).

**<u>Note on notation.</u>** I intend to use uppercase letters (*X*, *Y*, …) for random variables and lowercase letters (*x*, *y*, …) for particular (but unspecified) outcomes that the random variable can assume. Many books adopt this distinction. The distinction may seem pedantic at first, but it is important for clear thinking about random variables. The key to the distinction is that the random variable *X* is thought of as a *potential* outcome that occurs within a population of possible outcomes with probabilities assigned by the probability function, whereas the realization *x* is thought of as a particular *actual* outcome. After *x* has been observed, it no longer has a probability. Probabilities are reserved for potential outcomes.

> Ex: Suppose you toss a fair coin 10 times. If you define *X* to be the number of heads in the 10 tosses, then *X* is a RV.[2] *X* does not have a definite value until you have finished observing the tosses. Then it will become a definite value *x* (e.g., 3). *X* is random; *x* is not.

[1] An advisory! Most computer regression software automatically analyzes relationships between Y and X as though they were experiments. Researchers who have observational data (like rent and area) are often unaware of this and do not realize the issues involved.

[2] *X* has {0, 1, 2, …, 10} as its population of possible outcomes, and each of these outcomes has a probability given by $P(X=k)=\binom{10}{k}(1/2)^{10}$, *k*=0,1,2,…,10. These are the two things (outcomes and probabilities) required of a random variable.

Two types of random variables:

**Discrete**. For a discrete RV, the probability function assigns probabilities directly to the outcomes: $P(X = x_i) = f(x_i)$. So the probability of a range is computed by adding individual probabilities: $P(a < X < b) = \sum_{a < x_i < b} f(x_i)$.

Ex: Roll a fair die. Let $X$ = number of spots on side showing up. Then $P(X = i) = 1/6$ for $i$ = 1, 2, …, 6. The probability that the number of spots showing will be less than 3 is $P(X < 3) = P(X = 1) + P(X = 2) = 2/6$.

**Continuous**. For a continuous RV, the probability function assigns probabilities indirectly via a density function *f*(*x*). The probability of a range is computed by integration: $P(a < X < b) = \int_a^b f(x)dx$.

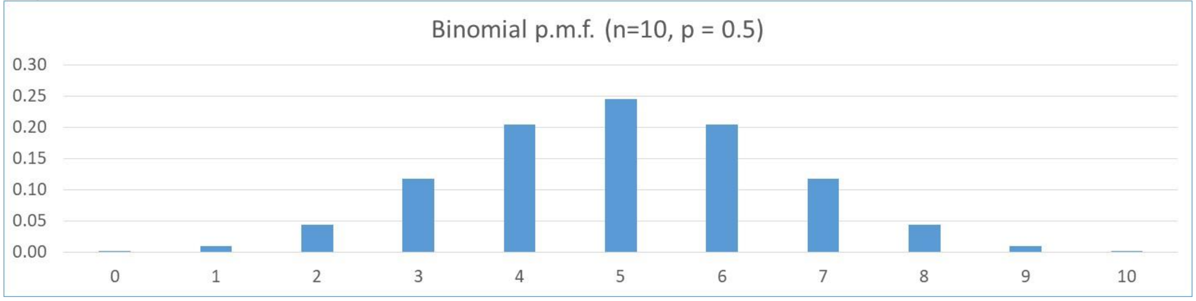
Ex: Suppose $X$ = the lifetime in hours of a light bulb, with density function $f(x) = \frac{1}{1000} e^{-x/1000}$.

Then the probability that the light bulb will last less than 1,000 hours is $P(X < 1000)$ = $\int_0^{1000} \frac{1}{1000} e^{-x/1000} dx = -e^{-x/1000} \Big|_{x=0}^{x=1000} = 1 - e^{-1}$.

For both discrete and continuous RVs, the probability function is called the **probability density function** or **p.d.f.** Sometimes the probability function for a discrete RV is called the **probability mass function** or **p.m.f.**

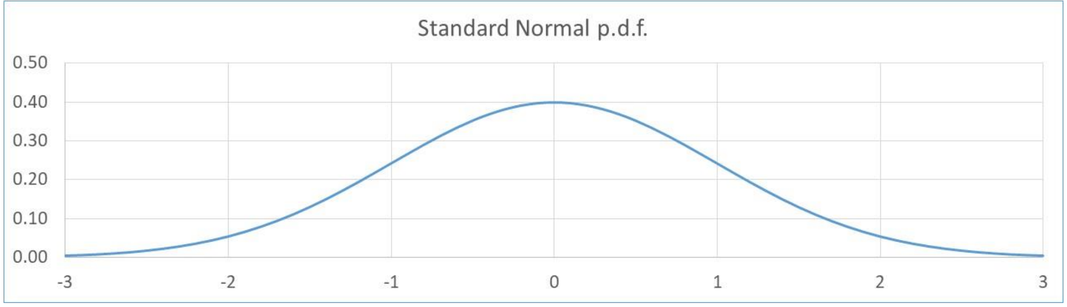
**Discussion.** The p.m.f. of a discrete RV is relatively easy to understand – the p.d.f. of a continuous RV not so much. The p.m.f. provides probabilities directly. For example, the p.m.f. of $X$ = number of heads in 10 independent tosses of a fair coin is $f(x) = \binom{10}{x}\left(\frac{1}{2}\right)^{10}$, $x = 0,1,...,10$. The graph of this p.m.f. is

Figure 5. Binomial p.m.f. (n=10, p=0.5).



The height of each bar is the probability of the value on the horizontal axis. The heights are all non-negative and sum to 1 – the necessary conditions for being a legitimate probability function. But what is the meaning of the p.d.f. of a continuous RV? For example, the p.d.f. of a standard normal RV is $f(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}$, $-\infty < x < \infty$. The graph of this p.d.f. is

Figure 6. Standard normal p.d.f.



Unlike the p.m.f., the value of the p.d.f. and the height of the plot are not the probability of the value on the horizontal axis. The value of the p.d.f. and the height of the plot at $x = 0$ are about 0.4. This is not the probability that $X = 0$.[3] Rather, the p.d.f. provides the degree of concentration of probability in the vicinity of $x$. The concentration of probability in an interval $(a,b)$ is the amount of probability per unit length = $P(a,b) / (b - a)$. For an interval $(x - \Delta x, x + \Delta x)$ around $x$, this is $P(x - \Delta x, x + \Delta x) / (2\Delta x) = \dfrac{\int_{x-\Delta x}^{x+\Delta x} f(u)du}{2\Delta x}$ . As the interval closes in on $x$, the concentration gets close to $\dfrac{\int_{x-\Delta x}^{x+\Delta x} f(u)du}{2\Delta x} \approx \dfrac{\int_{x-\Delta x}^{x+\Delta x} f(x)du}{2\Delta x} = \dfrac{f(x)\int_{x-\Delta x}^{x+\Delta x} du}{2\Delta x} = = \dfrac{f(x)[x+\Delta x-(x-\Delta x)]}{2\Delta x} = f(x)$. So the p.d.f. can be interpreted as the concentration of probability at each point $x$. Probability is highly concentrated where the p.d.f. is high.

**<u>Aside.</u>** The probability that a continuous RV $X$ *exactly* equals any given outcome $x$ is necessarily zero, as can be seen from the integration $P(x \le X \le x) = \int_x^x f(u)du = 0$. This may seem paradoxical, for is it not true that the range from $a$ to $b$ consists of all of the in-between values $x$? And if the probability of every one of the in-betweens is zero, does that not mean that when we add them all up we get zero for the probability of the whole range? Well, no. The catch is in "adding them up". There are infinitely many of them. Even worse: There are uncountably infinitely many. If there were only countably infinitely many, we could make an infinite list and define the sum to be the limit of the partial sums, as we do for infinite series: $\sum_{a<x<b} f(x) = \lim_{n\to\infty} \sum_{i=1}^{n} f(x_i)$ . But the partial sums are always 0, so their limit is 0, and that does not resolve the paradox. Instead we need a new method for "adding them up", and that method is

[3] Pr(X=0) = 0. Indeed, the probability that a continuous RV equals *any* single given value is zero. This sometimes perplexes students – but it should not. If I ask, what is the area between the standard normal p.d.f. and the horizontal axis in the plot? – you would answer, 1 – and you would be correct since the probability of the entire region must be 1. But suppose I ask, can you get that by adding up the area underneath each point on the curve and the corresponding $x$ on the axis? You might pause. You would reason, the area between the curve and the axis, for each point $x$, is the area of a line, which is zero. And if you add up a lot of zeroes, you still get zero – right? You might further say, There are infinitely many zeroes, so the answer would be ∞×0, which is undefined. Except, the answer is 1. You cannot add up the area of a bunch of lines to get the area of any two-dimensional figure; nor can you add up the probability of a bunch of points to get the probability of a range. It is the same paradox. Integration was invented to resolve the paradox and do the "adding up" properly to get the right answer.

integration. In order to make mathematics work, we must give up the intuitively reasonable idea that every possible outcome has a positive probability.

The same conceptual problem arises in thinking about the area of a plot of land. The plot can be covered from one side to the other by a series of infinitely thin lines, each one of which has 0 area. If each line has 0 area, then when you add them up, should you not get 0 for the area of the whole plot? This is exactly the same problem. And it has the same solution.

**End Aside.**

For both discrete and continuous RVs, the **cumulative distribution function** (**CDF**) gives the total probability less than or equal to *x* as *x* varies:

$$F(x) = P(-\infty < X < x) = \sum_{-\infty < x_i < b} f(x_i) \text{ (discrete)}$$

$$F(x) = P(-\infty < X < x) = \int_{-\infty}^{x} f(u)du \text{ (continuous)}$$

For continuous RVs, the CDF is the integral of the probability density function.

Ex: Suppose *X* = the lifetime in hours of a light bulb, with p.d.f. $f(x) = \frac{1}{1000}e^{-x/1000}$, for *x* > 0.

Then the CDF is $F(x) = P(-\infty < X < x) = \int_{-\infty}^{x} f(u)du = \int_{0}^{x} \frac{1}{1000}e^{-u/1000}du = -e^{-u/1000}\Big|_{x=0}^{x=1000}$

$= 1-e^{-x/1000}$, *x* > 0. Then the probability that the light bulb will last less than 1,000 hours is $F(1000) = 1-e^{-1000/1000} = 1-e^{-1}$.

<u>The most important examples of each type of RV:</u>

Discrete: **Binomial RV**.

The p.m.f. is $f(x) = \binom{n}{x}p^x(1-p)^{n-x}$, $x = 0,1,\ldots,n$.

*X* is the number of times a particular outcome occurs in *n* independent repetitions of the same experiment, in which *p* is the probability of the particular outcome in each repetition.

Ex: Toss a fair coin 10 times. Let *X* = number of heads in the 10 tosses. Then *X* is a binomial RV with a binomial probability distribution with *n* = 10 and *p* = ½. So

$P(X = x) = \binom{10}{x}(0.5)^x(1-0.5)^{10-x}$. E.g., $P(X = 3) = \binom{10}{3}(0.5)^3(1-0.5)^{10-3}$

Continuous: **Normal RV**

The p.d.f. is $f(x) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$, $-\infty < x < +\infty$. **$\mu$** is the mean of the distribution, **$\sigma^2$** is the variance.

Ex: If *X* is a standard normal RV ($\mu = 0$, $\sigma^2 = 1$), then $P(-1 < X < +1) = \int_{-1}^{+1}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}dx \approx$ 0.6827.

There is no closed-form elementary function that is the indefinite integral of $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$. So the CDF cannot be written in closed-form. This does not mean that there is no CDF. It just means that its value must be approximated by numerical algorithms.

Mathematical expectation.

| | | |
|---|---|---|
| $E(X) =$ | $\sum_{all\ x_i} x_i f(x_i)$ | Discrete RV |
| | $\int_{-\infty}^{+\infty} xf(x)dx$ | Continuous RV |

This is the **mean** or **expected value** of *X* – denoted $\mu$. The mean can be thought of as the average value (arithmetic mean) of a very large number of independent draws from the distribution of a random variable.

The mean of the random variable $(X - E(X))^2$ is called the **variance** of *X* – denoted $\sigma^2$:

| | | |
|---|---|---|
| $E\{(X - E(X))^2\} =$ | $\sum_{all\ x_i} (x_i - E(X))^2 f(x_i)$ | Discrete RV |
| | $\int_{-\infty}^{+\infty} (x - E(X))^2 f(x)dx$ | Continuous RV |

The square root of the variance is called the **standard deviation** – denoted $\sigma$.

The mean and variance are the two most important characteristics of most RVs. The mean is a good measure of centrality because the negative deviations from the value *c* exactly average out against the positive deviations from *c* when *c* is chosen to be the mean – and for no other value of *c*. The variance is a good measure of the spread of the values of the RV around its mean, because the variance is the average magnitude of those (squared) deviations from the mean.

Ex: Roll a fair die. Let *X* = number of spots on side showing up. Then $P(X = i) = 1/6$ for *i* = 1, 2, …, 6. Then $E(X)$ = $\sum_{x=1}^{6} xP(X = x)$ = $\sum_{x=1}^{6} x\frac{1}{6}$ = 3.5.

$Var(X)$ = $\sum_{x=1}^{6} (x - 3.5)^2 P(X = x)$ = $\sum_{x=1}^{6} (x - 3.5)^2 \frac{1}{6}$ = 35/12.

Ex: Suppose *X* = the lifetime in hours of a light bulb, with density function $f(x) = \frac{1}{1000} e^{-x/1000}$.

Then $E(X)$ = $\int_0^{\infty} x\frac{1}{1000} e^{-x/1000} dx$ = (integrate by parts) $(-x\, e^{-x/1000} - 1000e^{-x/1000})\Big|_{x=0}^{x=\infty}$ = 1,000.

$Var(X)$ = $\int_0^{\infty} (x - 1000)^2 \frac{1}{1000} e^{-x/1000} dx$ = $\int_0^{\infty} x^2 \frac{1}{1000} e^{-x/1000} dx - (1000)^2$ = (integrate successively by parts) $(-x^2\, e^{-x/1000} - 2000x\, e^{-x/1000} - 2000000e^{-x/1000})\Big|_{x=0}^{x=\infty}$ - 1,000,000 = 2,000,000 – 1,000,000 = 1,000,000.

Some properties of expectation:

For all RVs –

1) $E(X-\mu)=0$. In words: The average deviation of a RV from its own mean is zero.
2) If $X = c$ with probability 1 ($X$ is a constant), then $E(X)=c$. In words: You can expect a constant RV to be that constant.
3) If $c$ is a constant, then $E\{cg(X)\}=cE\{g(X)\}$.
   Ex: $E\{2\sqrt{x}\}=2E\{\sqrt{x}\}$
4) $E\{u(X)+v(X)\}=E\{u(X)\}+E\{v(X)\}$. In words, the mean of a sum is the sum of the means (but the mean of a product is not usually the product of the means).
   Ex: $E\{X^2-X\}=E\{X^2\}-E\{X\}$
5) $\sigma^2=E\{(X-\mu)^2\}=E\{X^2\}-\mu^2$. This is a convenient formula for computing a variance.
6) If $X = c$ with probability 1 ($X$ is a constant), then $Var(X)=0$. In words, a constant has no variability.
7) If $a$ and $b$ are constants, then $Var(a+bX)=b^2Var(X)$.
   Ex: $Var\{2-3X\}=9Var\{X\}$

# Section 3. Multiple Random Variables

The concept of random variable (RV) can be extended to cover a process of observation or experimentation that produces two or more observations or outcomes. There are three ways that multiple outcomes can happen:

- The process can be repeated (e.g., two tosses of the same coin); or
- The process can produce single measurements on different quantities (e.g., the toss of a coin and the throw of a die); or
- Both (e.g., two tosses of a coin and two throws of a die).

The best way to extend the concept of RV to multiple outcomes is to use a single model that preserves the possible correlations among the RVs but allows the individual RVs to be extracted if necessary. The concept of jointly distributed RVs does this. To implement this, we need to define a possible outcome to be a vector of observations $(x_1, x_2, ..., x_n)$.

The set of all possible vector outcomes for the observation or experiment is the **population**. A collection of observation vectors or experimental outcome vectors is a **sample**.[4]

> Ex: If you collect data on individual companies, say their stock price *y*, earnings *x*, and total assets *z*, then the collection of all triples (*x*, *y*, *z*) – one triple for each company – is your population. If you take a subset of 40 companies, then their triples of price, earnings, and assets $\{(x_1, y_1, z_1), ..., (x_{40}, y_{40}, z_{40})\}$ are your sample.

A **random variable (random vector)** $(X_1, X_2, ..., X_n)$ is a population of outcomes $\{(x_1, x_2, ..., x_n)$'s$\}$ and a probability function $f(x_1, x_2, ..., x_n)$ that assigns probabilities to those outcomes.

I will highlight the properties of joint distributions for two random variables. The extension to more than two is straightforward.

<u>Two types of joint random variables:</u>

**Discrete**. For a discrete joint RV, the probability function (p.m.f.) assigns probabilities directly to the outcomes: $P(X = x_i, Y = y_j) = f(x_i, y_j)$.[5] So the probability of a range is computed by adding individual probabilities: $P(a < X < b, c < Y < d) = \sum_{a < x_i < b} \sum_{c < y_j < d} f(x_i, y_j)$

> Ex: Suppose $f(x_i, y_j) = \frac{x_i y_j}{18} + \frac{x_i}{12}$, for $x_1 = 1, x_2 = 2, y_1 = 1, y_2 = 2$. To qualify as a p.m.f, the function must have $f(x_i, y_j) \geq 0$ and $\sum_{i=1}^{2} \sum_{j=1}^{2} f(x_i, y_j) = 1$ [Why?]. [Verify that these conditions are satisfied.]

[4] When each RV is a repetition of one experiment or observation, a single vector $(x_1, x_2, ..., x_n)$ may be considered to be a sample itself.

[5] This means the probability that $X = x_i$ <u>AND</u> $Y = y_j$ simultaneously.

**Continuous**. For a continuous RV, the probability function (p.d.f.) assigns probabilities indirectly via a joint density function. So the probability of a range is computed by multiple integration: $P(a < X < b, c < Y < d) = \int_a^b \int_c^d f(x,y)\,dy\,dx$ .

Ex: Suppose $f(x,y) = 4xy$ if $0 < x < 1, 0 < y < 1$. To qualify as a p.d.f., the function must have $f(x,y) \geq 0$ and $\int_0^1 \int_0^1 f(x,y)\,dy\,dx = 1$. [Why? Verify that these conditions are satisfied.]

The joint distribution of (*X*,*Y*) determines the individual distributions of *X* and of *Y* and also the nature of the dependence between them. The individual distribution of *X* can be obtained by “summing out” *Y* (discrete case) or by “integrating out” *Y* (continuous case). This process of summing/integrating out the joint density to get the density of the individual variable is given below in the definition of marginal density. The nature of the dependence between *X* and *Y* is revealed by the conditional distributions of *X* and *Y*, given the value of the other variable.

The **marginal density functions** of *X* and *Y* can be obtained by

Discrete case: $f_X(x_i) = \sum_{all\ y_j} f(x_i, y_j)$ , $f_Y(y_j) = \sum_{all\ x_i} f(x_i, y_j)$

Continuous case: $f_X(x) = \int_{-\infty}^{+\infty} f(x,y)\,dy$ , $f_Y(y) = \int_{-\infty}^{+\infty} f(x,y)\,dx$ .

This is intuitive in the discrete case, for the densities are p.m.f.’s and so are probabilities. Thus, $f_X(x_1) = P(X = x_1) = P(X = x_1 \text{ and } Y = \text{ anything}) = P(X = x_1, -\infty < Y < +\infty) =$ $\sum_{-\infty<y_j}^{y_j<+\infty} P(X = x_1, Y = y_j) = \sum_{-\infty<y_j}^{y_j<+\infty} f(x_1, y_j)$. In words, you pick a particular value $x_1$ of *X*, then look at all the ways it can pair up with values for *Y*, and add the probabilities of all those ways. Do that for each possible $x_1$.

In the continuous case, the densities are p.d.f.’s and so are concentrations of probability. To get the concentration of *X* probability at *X* = *x*, you look at the concentration near *x*: For an interval (*x* – Δ*x*, *x* + Δ*x*) around *x*, this is P(*x* – Δ*x* < *X* < *x* + Δ*x*) / (2Δ*x*) = P(*x* – Δ*x* < *X* < *x* + Δ*x* and Y is anything) / (2Δ*x*) = P(*x* – Δ*x* < *X* < *x* + Δ*x*, -∞ < *Y* < +∞) / (2Δ*x*) = $\frac{\int_{-\infty<y}^{y<+\infty} \int_{x-\Delta x}^{x+\Delta x} f(u,y)\,du\,dy}{2\Delta x}$ . As the interval closes in on *x*, the concentration gets close to $\frac{\int_{-\infty<y}^{y<+\infty} \int_{x-\Delta x}^{x+\Delta x} f(x,y)\,du\,dy}{2\Delta x}$ =

$$\frac{\int_{-\infty<y}^{y<+\infty} f(x,y)\left[\int_{x-\Delta x}^{x+\Delta x} du\right]dy}{2\Delta x} = \frac{\int_{-\infty<y}^{y<+\infty} f(x,y)\left[(x+\Delta x)-(x-\Delta x)\right]dy}{2\Delta x} = \int_{-\infty<y}^{y<+\infty} f(x,y)dy .$$

Ex: (Discrete) Suppose $f(x_i, y_j) = \frac{x_i y_j}{18} + \frac{x_i}{12}$, for $x_1 = 1, x_2 = 2, y_1 = 1, y_2 = 2$. Then the marginal density of *X* is $f_X(x_1) = \sum_{all\ y_j} f(x_1, y_j) =$ (“sum out *y*”) $f(x_1, y_1) + f(x_1, y_2)$ = (1/18 +

1/12) + (2/18 + 1/12) = 1/3, and $f_X(x_2) = \sum_{all\ y_j} f(x_2, y_j) = f(x_2, y_1) + f(x_2, y_2)$ = (2/18 + 2/12) + (4/18 + 2/12) = 2/3.

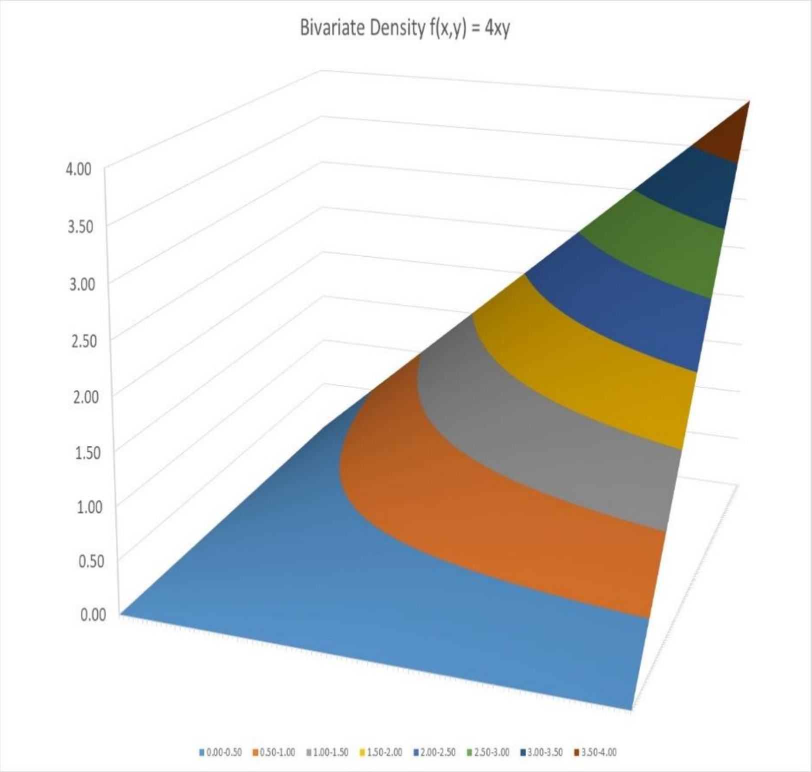
And the marginal density of *Y* is $f_Y(y_1) = \sum_{all\ x_i} f(x_i, y_1)$ = ("sum out *x*") $f(x_1, y_1) + f(x_2, y_1)$ = (1/18 + 1/12) + (2/18 + 2/12) = 5/12. And $f_Y(y_2) = \sum_{all\ x_i} f(x_i, y_2)$ = $f(x_1, y_2) + f(x_2, y_2)$ = (2/18 + 1/12) + (4/18 + 2/12) = 7/12.

Notice that both computations are legitimate p.m.f.'s, in that they are nonnegative and sum to 1 (1/3 + 2/3 = 1 and 5/12 + 7/12 = 1).

Ex: (Continuous) Suppose $f(x, y) = 4xy$ if $0 < x < 1, 0 < y < 1$. Then $f_X(x) = \int_{-\infty}^{+\infty} f(x, y)\, dy$ = ("integrate out *y*") $\int_0^1 4xy\, dy = 2x$ , for each *x*, $0 < x < 1$. [Verify that this is a valid density function: it is nonnegative and integrates to 1.]

And $f_Y(y) = \int_{-\infty}^{+\infty} f(x, y)\, dx$ = ("integrate out *x*") $\int_0^1 4xy\, dx = 2y$ , if $0 < y < 1$. [Verify that this is a valid density function: it is nonnegative and integrates to 1.]

Figure 7. Bivariate p.d.f. *f*(*x*,*y*) = 4*xy* (the axes not drawn to scale).



The **conditional density functions** of *X* given *Y* = *y* and of *Y* given *X* = *x* are (for both discrete and continuous cases): $f_{X|Y=y}(x) = \dfrac{f(x, y)}{f_Y(y)}$, $f_{Y|X=x}(y) = \dfrac{f(x, y)}{f_X(x)}$ .

Note that in the discrete case, the conditional densities (p.m.f.'s) are just conditional probabilities: $P(X = x \mid Y = y) = \dfrac{P(X = x \text{ and } Y = y)}{P(Y = y)}$ and

$P(Y = y \mid X = x) = \dfrac{P(X = x \text{ and } Y = y)}{P(X = x)}$.

In the continuous case, the conditional densities are relative concentrations of probability. To get the conditional concentration of *X* probability at *X* = *x*, relative to a given *Y* = *y*, you look at the *X* concentration near *x*, given *Y* being near *y*: For a rectangle (*x* – Δ*x* < *x* < *x* + Δ*x*, *y* – Δ*y* < *y* < *y* + Δ*y*) around *x* and *y*, this conditional concentration for *X* given *Y* is P(*x* – Δ*x* < *X* < *x* + Δ*x* | *y* – Δ*y* < *Y* < *y* + Δ*y*)/ (2Δ*x*) = P(*x* – Δ*x* < *X* < *x* + Δ*x* and *y* – Δ*y* < *Y* < *y* + Δ*y*) / (2Δ*x P(y* – Δ*y* < *Y* < *y* + Δ*y*)) = $\dfrac{\int_{y-\Delta y}^{y+\Delta y}\int_{x-\Delta x}^{x+\Delta x} f(u,v)\,du\,dv}{2\Delta x \int_{y-\Delta y}^{y+\Delta y} f_Y(v)\,dv}$ . As the *X* interval closes in on *x*, this ratio gets close to

$$\frac{\int_{y-\Delta y}^{y+\Delta y}\int_{x-\Delta x}^{x+\Delta x} f(x,v)\,du\,dv}{2\Delta x \int_{y-\Delta y}^{y+\Delta y} f_Y(v)\,dv} = \frac{\int_{y-\Delta y}^{y+\Delta y} f(x,v)\int_{x-\Delta x}^{x+\Delta x} du\,dv}{2\Delta x \int_{y-\Delta y}^{y+\Delta y} f_Y(v)\,dv} = \frac{\int_{y-\Delta y}^{y+\Delta y} f(x,v)[(x+\Delta x)-(x-\Delta x)]\,dv}{2\Delta x \int_{y-\Delta y}^{y+\Delta y} f_Y(v)\,dv} =$$

$\dfrac{\int_{y-\Delta y}^{y+\Delta y} f(x,v)\,dv}{\int_{y-\Delta y}^{y+\Delta y} f_Y(v)\,dv}$. As the remaining *Y* interval closes in on *y*, this ratio gets close to

$$\frac{\int_{y-\Delta y}^{y+\Delta y} f(x,y)\,dv}{\int_{y-\Delta y}^{y+\Delta y} f_Y(y)\,dv} = \frac{f(x,y)\int_{y-\Delta y}^{y+\Delta y} dv}{f_Y(y)\int_{y-\Delta y}^{y+\Delta y} dv} = \frac{f(x,y)[(y+\Delta y)-(y-\Delta y)]}{f_Y(y)[(y+\Delta y)-(y-\Delta y)]} = \frac{f(x,y)}{f_Y(y)}.$$

Ex: Suppose $f(x_i, y_j) = \dfrac{x_i y_j}{18} + \dfrac{x_i}{12}$, for $x_1 = 1, x_2 = 2, y_1 = 1, y_2 = 2$.

Then the conditional distribution of *X* given *Y* = 1 is $f_{X|Y=y_1}(x_1) = \dfrac{f(x_1, y_1)}{f_Y(y_1)}$ = $\dfrac{1/18 + 1/12}{5/12}$ = 1/3 and $f_{X|Y=y_1}(x_2) = \dfrac{f(x_2, y_1)}{f_Y(y_1)}$ = $\dfrac{2/18 + 2/12}{5/12}$ =2/3.

And given *Y* = 2 is $f_{X|Y=y_2}(x_1) = \dfrac{f(x_1, y_2)}{f_Y(y_2)}$ = $\dfrac{2/18 + 1/12}{7/12}$ = 1/3 and $f_{X|Y=y_2}(x_2) = \dfrac{f(x_2, y_2)}{f_Y(y_2)}$ = $\dfrac{4/18 + 2/12}{7/12}$ =2/3.

Also, the conditional distribution of *Y* given *X* = 1 is $f_{Y|X=x_1}(y_1) = \dfrac{f(x_1, y_1)}{f_X(x_1)}$ = $\dfrac{1/18 + 1/12}{1/3}$ = 5/12 and $f_{Y|X=x_1}(y_2) = \dfrac{f(x_1, y_2)}{f_X(x_1)}$ = $\dfrac{2/18 + 1/12}{1/3}$ =7/12.

And given $X$ = 2 is $f_{Y|X=x_2}(y_1) = \frac{f(x_2, y_1)}{f_X(x_2)}$ = $\frac{2/18 + 2/12}{2/3}$ = 5/12 and

$f_{Y|X=x_2}(y_2) = \frac{f(x_2, y_2)}{f_X(x_2)}$ = $\frac{4/18 + 2/12}{2/3}$ =7/12.

One very important kind of relationship between $X$ and $Y$ is *independence*. $X$ and $Y$ are **independent** if their joint density is the product of their marginal densities: $f(x, y) = f_X(x) f_Y(y)$ for all $x$ and $y$ (both discrete and continuous cases).

- Note that independence of RVs is just independence for all of their probabilities: $P(X = x \text{ and } Y = y) = P(X = x)P(Y = y)$ for all $x$ and $y$ (discrete case).
- Note also that independence of RVs says that the conditional densities are the same as the marginal densities: $f_{X|Y=y}(x) = f_X(x)$ and $f_{Y|X=x}(y) = f_Y(y)$. This is true because independence says $f(x, y) = f_X(x) f_Y(y)$. Dividing both sides by $f_Y(y)$ yields $\frac{f(x, y)}{f_Y(y)} = f_X(x)$, or $f_{X|Y=y}(x) = f_X(x)$ Intuitively, this says that knowing that $Y = y$ does not affect the probability that $X = x$, and knowing that $X = x$ does not affect the probability that $Y = y$. That is, the value of $Y$ is irrelevant to probabilities for $X$ and vice-versa.

  Ex: (The preceding discrete example.) Suppose $f(x_i, y_j) = \frac{x_i y_j}{18} + \frac{x_i}{12}$, for $x_1 = 1, x_2 = 2, y_1 = 1, y_2 = 2$. Then $X$ and $Y$ are independent. For example, $f(x_1, y_1)$ = 1/18 + 1/12 = 5/36, which equals the product of $f_X(x_1)$ = 1/3 and $f_Y(y_1)$ = 5/12. [Verify that $f(x_i, y_j) = f_X(x_i) f_Y(y_j)$ for all other combinations of $x_i$ and $y_j$.]

  Ex: (Continuous case.) Suppose $f(x, y) = 4xy$ if $0 < x < 1, 0 < y < 1$. Then $X$ and $Y$ are independent because $4xy = (2x)(2y) = f_X(x) f_Y(y)$ for all $0 < x < 1, 0 < y < 1$.

Mathematical expectation for a function of jointly distributed RVs is a natural extension of expected values for single RVs. Although we have more than one variable in the joint case, mathematical expectation is always for just one RV. E.g., we can compute the expected value of $XY$ ($X$ times $Y$) because $XY$ is just one RV, although it is computed from two RVs. We do not compute the expected value of ($X$,$Y$).[6]

The **expected value of** $g(X, Y)$ is

Discrete case: $E[g(X, Y)] = \sum_{\text{all } x_i} \sum_{\text{all } y_j} g(x_i, y_j) f(x_i, y_j)$

[6] If we do, then we mean (E($X$),E($Y$)) – i.e., the two separate expectations, rather than some exotic two-dimensional concept.

Ex: Suppose $f(x_i, y_j) = \frac{x_i y_j}{18} + \frac{x_i}{12}$, for $x_1 = 1, x_2 = 2, y_1 = 1, y_2 = 2$ and $g(X,Y) = XY$. Then $E[g(X,Y)] = \sum_{all\ x_i} \sum_{all\ y_j} x_i y_j \left( \frac{x_i y_j}{18} + \frac{x_i}{12} \right) = \sum_{i=1}^{2} \sum_{j=1}^{2} \left( \frac{x_i^2 y_j^2}{18} + \frac{x_i^2 y_j}{12} \right)$ = (1/18 + 1/12) + (4/18 + 2/12) + (4/18 +4/12) + (16/18 + 8/12) = 95/36.

Continuous case: $E[g(X,Y)] = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} g(x,y) f(x,y)\, dx\, dy$

Ex: Suppose $f(x,y) = 4xy$ if $0 < x < 1, 0 < y < 1$, and $g(X,Y) = XY$. Then $E[g(X,Y)] = \int_0^1 \int_0^1 xy \cdot 4xy\, dx\, dy = \int_0^1 \int_0^1 4x^2 y^2\, dx\, dy$ = 4/9.

The expected value for conditional distributions is defined analogously. Here are the conditional mean and variance.

The conditional mean of *Y* given *X* = *x* is

- Discrete case: $E(Y \mid X = x_i) = \sum_{all\ y_j} y_j f_{Y|X=x_i}(y_j)$
- Continuous case: $E(Y \mid X = x) = \int_{-\infty}^{+\infty} y f_{Y|X=x}(y)\, dy$

The conditional variance of *Y* given *X* = *x* is

- Discrete case: $Var(Y \mid X = x_i) = \sum_{all\ y_j} \{y_j - E(Y \mid X = x_i)\}^2 f_{Y|X=x_i}(y_j)$
- Continuous case: $Var(Y \mid X = x) = \int_{-\infty}^{+\infty} \{y - E(Y \mid X = x)\}^2 f_{Y|X=x}(y)\, dy$

Analogously, the conditional mean of *X* given *Y* = *y* is

- Discrete case: $E(X \mid Y = y_j) = \sum_{all\ x_i} x_i f_{X|Y=y_j}(x_i)$
- Continuous case: $E(X \mid Y = y) = \int_{-\infty}^{+\infty} x f_{X|Y=y}(x)\, dx$

The conditional variance of *X* given *Y* = *y* is

- Discrete case: $Var(X \mid Y = y_j) = \sum_{all\ x_i} \{x_i - E(X \mid Y = y_j)\}^2 f_{X|Y=y_j}(x_i)$
- Continuous case: $Var(X \mid Y = y) = \int_{-\infty}^{+\infty} \{x - E(X \mid Y = y)\}^2 f_{X|Y=y}(x)\, dx$

The preceding concepts play major roles in regression analysis because regression seeks to estimate the conditional mean of *Y* given the value of *X*. In fact, the regression equation is the conditional mean of *Y* given *X* = *x*, as *x* varies. The conditional variance of *Y* given *X* = *x* provides an assessment of the accuracy of the regression because it measures how closely the *Y* values cluster around the regression equation.

Ex: Suppose that *Y* is monthly rent ($) and *X* is area (square feet) of apartments. Suppose that the conditional mean of *Y* given *X* = *x* is 160 + 0.50 *x*. This is the regression equation. In this case, it is a linear function of *x*, as *x* varies. (The conditional mean need not be linear, in general.) It says that the mean rent of 1000 square foot apartments (for example) is $660, and that the mean rent increases by 50 cents for each additional square foot of area. Suppose that the conditional variance of *Y* given *X* = *x* is $0.04x^2$. Then the standard deviation of rents of 400 square foot apartments is $80, and the standard deviation of rents of 1000 square foot apartments is $200. So in this example, rents are more variable, the larger the apartment.

Among the most commonly used measures of the strength of the relationship between *X* and *Y* are covariance and correlation. Strictly speaking, they measure only a certain type of relationship – linear relationship – the extent to which the probabilities for *X* and *Y* are concentrated around a line $Y = a + bX$ or $X = c + dY$. But if the issue is only whether *X* and *Y* are related or not – whether linearly or otherwise – then covariance and correlation are *often* sufficient to establish this. The reason is that a nonlinear relationship between *X* and *Y* will usually have some portion that is linear enough to affect the linear correlation measure.

The **covariance** between *X* and *Y* is $Cov(X,Y) = E(\{X - E(X)\}\{Y - E(Y)\}) = E(XY) - E(X)E(Y)$, denoted by $\sigma_{xy}$.

The **correlation coefficient** between *X* and *Y* is $Corr(X,Y) = \dfrac{Cov(X,Y)}{\sigma_X \sigma_Y}$, denoted by $\rho_{xy}$.

The correlation coefficient is just a scaled version of the covariance that makes it unit-less and keeps it between -1 and +1.

Covariance is actually a rather intuitive measure of the strength of relationship between *X* and *Y*. For example, consider rent (*Y*) and area (*X*) of apartments. If an apartment is large, then its area exceeds the mean: *X* – E(*X*) > 0. Most, but not all, large apartments also have rents larger than average: *Y* – E(*Y*) > 0. Similarly, if an apartment is small, then its area is less than the mean: *X* – E(*X*) < 0. Most, but not all, small apartments also have rents smaller than average: *Y* – E(*Y*) < 0. For both large and small apartments, most of the *X* and *Y* deviations have the same sign, so their product deviations [*X* – E(*X*)][*Y* – E(*Y*)] > 0. The stronger the relationship between size and rent, the more apartments have positive product deviations. So the average product deviation (the expected value) E{[*X* – E(*X*)][*Y* – E(*Y*)]} > 0. The more mis-matches there are of signs – the more large apartments with small rents and the more small apartments with large rents – the more negative product deviations there are and the more they offset the positive products, making the covariance smaller. Variables *X* and *Y* that tend to run in opposite directions, like (perhaps) apartment age and rent, tend to have negative product deviations and negative covariances. Variables that have little relationship with each other tend to have a random mix of positive and negative product deviations, which tend to cancel out each other, leaving a covariance around zero. So the sign of covariance indicates the direction of the relationship between *X* and *Y*, the larger the magnitude of the covariance, the stronger the relationship, and correlation is merely covariance adjusted to a universal scale.

Ex: Suppose $f(x,y) = 4xy$ if $0 < x < 1, 0 < y < 1$. Then $E(XY)$ = 4/9 (this was shown in a previous example), and $E(X) = \int_0^1 x \cdot 2x\,dx$ = 2/3, and $E(Y) = \int_0^1 y \cdot 2y\,dy$ = 2/3. So $Cov(X,Y) = E(XY) - E(X)E(Y)$ = 4/9 – (2/3)(2/3) = 0. So *X* and *Y* are uncorrelated (which also follows from the fact that *X* and *Y* are independent.)

Some properties:

- $Cov(X,X) = Var(X)$
- If $X$ and $Y$ are independent, then $E(g(X)h(Y)) = E(g(X)) \cdot E(h(Y))$. As a consequence, if $X$ and $Y$ are independent, then $E(XY) = E(X)E(Y)$, so $Cov(X,Y) = 0$ and $Corr(X,Y) = 0$.[7]
- Always: $-1 \le Corr(X,Y) \le +1$.
- $|Corr(X,Y)| = 1$ if and only if $Y = a + bX$ for some $a$ and $b \ne 0$.

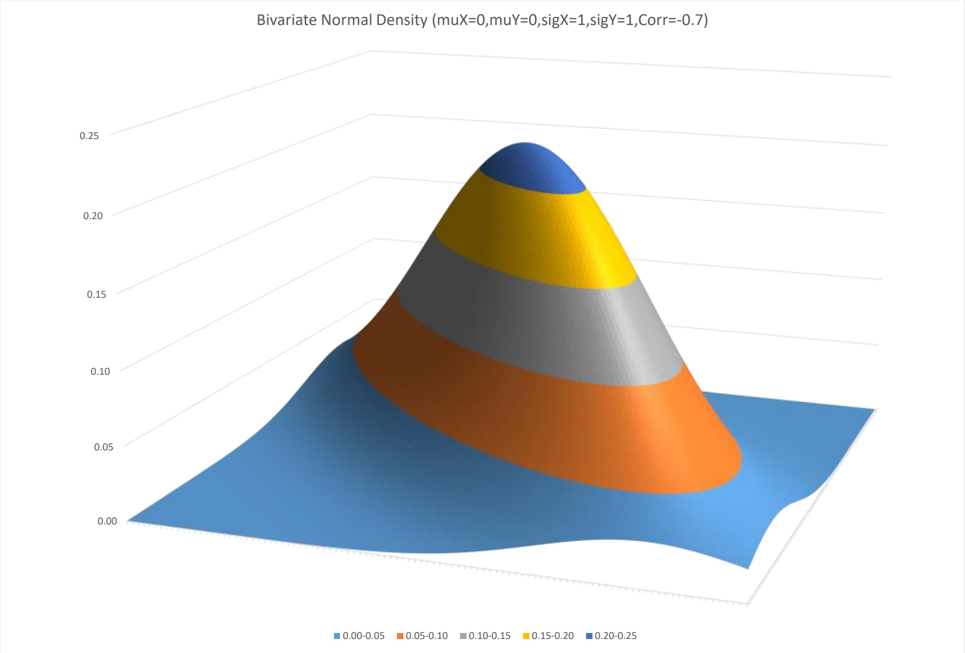
## The Bivariate Normal Distribution

The multivariate normal distribution is the most important multivariate distribution in statistics. The reason for its importance is the large role that it plays in the analysis of models for relationships among variables, like regression. Here, I will present the bivariate normal distribution. Jointly distributed random variables (*X*,*Y*) have a **bivariate normal distribution** if their joint probability density function is

$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}\exp\left(-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_x)^2}{\sigma_x{}^2} - \frac{2\rho(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y} + \frac{(y-\mu_y)^2}{\sigma_y{}^2}\right]\right)$$

for all (*x*,*y*). The bivariate normal distribution has 5 parameters: $\mu_x$ is the mean of *X*; $\sigma_x{}^2$ is the variance of *X*; $\mu_y$ is the mean of *Y*; $\sigma_y{}^2$ is the variance of *Y*; $\rho$ is the correlation coefficient between *X* and *Y*. Figure 8 shows an example of a bivariate normal density.

Figure 8. Bivariate normal density with ($\mu_x, \mu_y, \sigma_x, \sigma_y, \rho$) = (0,0,1,1,-0.7).



As can be seen in Figure 8, the locus of points where the density is constant is an ellipse. That is, if you cut the density plot with a plane that is parallel to the *x*,*y* base, you get an ellipse

[7] The converse is not true: Examples can be given in which the correlation is zero but *X* and *Y* are dependent.

(see the boundaries of the colored bands). The closer to zero the correlation, the more circular the ellipse. The more circular the ellipse, the weaker the relationship between *X* and *Y*. To see this clearly, suppose that the correlation $\rho$ = 0. Then upon substitution of $\rho$ = 0 into *f*(*x*,*y*), the bivariate normal density becomes

$$f(x,y)=\frac{1}{2\pi\sigma_x\sigma_y}\exp\left(-\frac{1}{2}\left[\frac{(x-\mu_x)^2}{\sigma_x{}^2}+\frac{(y-\mu_y)^2}{\sigma_y{}^2}\right]\right) =$$

$$\frac{1}{\sqrt{2\pi}\sigma_x}\exp\left(-\frac{1}{2}\left[\frac{(x-\mu_x)^2}{\sigma_x{}^2}\right]\right)\times\frac{1}{\sqrt{2\pi}\sigma_y}\exp\left(-\frac{1}{2}\left[\frac{(y-\mu_y)^2}{\sigma_y{}^2}\right]\right),$$

which is the product of a normal density of *X* with a normal density of *Y*. Therefore, *X* and *Y* are independent and hence unrelated to each other.

I have asserted, but not shown, that if (*X*,*Y*) is bivariate normal, then *X* is N($\mu_x,\sigma_x{}^2$) and *Y* is N($\mu_y,\sigma_y{}^2$). To satisfy yourself that this is true, you can get the density of *X* by "integrating out" *Y* in *f*(*x*,*y*) and the density of *Y* by "integrating out" *X* in *f*(*x*,*y*). This is tedious and involves a trick:

$$f_X(x)=\int_{-\infty}^{+\infty}\frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}\exp\left(-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_x)^2}{\sigma_x{}^2}-\frac{2\rho(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y}+\frac{(y-\mu_y)^2}{\sigma_y{}^2}\right]\right)dy$$

$$=\int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi}\sigma_x}\exp\left(-\frac{1}{2}\frac{(x-\mu_x)^2}{\sigma_x{}^2}\right)\times\frac{1}{\sqrt{2\pi}\sigma_y\sqrt{1-\rho^2}}\exp\left(-\frac{1}{2}\frac{(y-\mu_y-\rho\frac{\sigma_y}{\sigma_x}(x-\mu_x))^2}{\sigma_y{}^2(1-\rho^2)}\right)dy$$

The tedious part is the algebra to show that the integrands of the preceding two expressions are equal. The trick is that the integrand is now the product of two normal densities: the first density is N($\mu_x,\sigma_x{}^2$); the second is normal with mean $\mu_y+\rho\frac{\sigma_y}{\sigma_x}(x-\mu_x)$ and variance $\sigma_y{}^2(1-\rho^2)$. The first density does not involve *y* and so may be brought outside the integral. The integrand is therefore a normal density being integrated over its entire range – which yields 1:

$$f_X(x)=\frac{1}{\sqrt{2\pi}\sigma_x}\exp\left(-\frac{1}{2}\frac{(x-\mu_x)^2}{\sigma_x{}^2}\right)\int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi}\sigma_y\sqrt{1-\rho^2}}\exp\left(-\frac{1}{2}\frac{(y-\mu_y-\rho\frac{\sigma_y}{\sigma_x}(x-\mu_x))^2}{\sigma_y{}^2(1-\rho^2)}\right)dy\ =$$

$\frac{1}{\sqrt{2\pi}\sigma_x}\exp\left(-\frac{1}{2}\frac{(x-\mu_x)^2}{\sigma_x{}^2}\right)\times 1$, which says that *X* is N($\mu_x,\sigma_x{}^2$). Analogously, *Y* is N($\mu_y,\sigma_y{}^2$).

This argument also provides us the conditional density of *Y* given *X* = *x* and of *X* given *Y* = *y* as follows: Since the rewritten integrand is still *f*(*x*,*y*), we have

$$f_Y(y \mid X=x)=\frac{f(x,y)}{f_X(x)}=\frac{\dfrac{1}{\sqrt{2\pi}\sigma_x}\exp\left(-\dfrac{1}{2}\dfrac{(x-\mu_x)^2}{\sigma_x{}^2}\right)\times\dfrac{1}{\sqrt{2\pi}\sigma_y\sqrt{1-\rho^2}}\exp\left(-\dfrac{1}{2}\dfrac{(y-\mu_y-\rho\dfrac{\sigma_y}{\sigma_x}(x-\mu_x))^2}{\sigma_y{}^2(1-\rho^2)}\right)}{\dfrac{1}{\sqrt{2\pi}\sigma_x}\exp\left(-\dfrac{1}{2}\dfrac{(x-\mu_x)^2}{\sigma_x{}^2}\right)}$$

$$=\ \frac{1}{\sqrt{2\pi}\sigma_y\sqrt{1-\rho^2}}\exp\left(-\frac{1}{2}\frac{(y-\mu_y-\rho\dfrac{\sigma_y}{\sigma_x}(x-\mu_x))^2}{\sigma_y{}^2(1-\rho^2)}\right).$$

Therefore, the conditional distribution of *Y* given *X* = *x* is N( $\mu_y+\rho\frac{\sigma_y}{\sigma_x}(x-\mu_x)$, $\sigma_y{}^2(1-\rho^2)$ ). Analogously, the conditional distribution of *X* given *Y* = *y* is N( $\mu_x+\rho\frac{\sigma_x}{\sigma_y}(y-\mu_y)$ , $\sigma_x{}^2(1-\rho^2)$ ). An *extremely* important application of this result occurs in regression: The major purpose of regression is to estimate the mean of *Y* given the value *x* for *X*. Under the specifications (assumptions) of the standard regression model, the conditional distribution of *Y* given *X* = *x* is normal. That conditional mean is $\mu_y+\rho\frac{\sigma_y}{\sigma_x}(x-\mu_x)$. *This is the true regression equation*. Notice that it is a linear function of *x*, with slope = $\rho\frac{\sigma_y}{\sigma_x}$ and intercept = $\mu_y-\rho\frac{\sigma_y}{\sigma_x}\mu_x$. The sample estimate of the regression equation is found by substituting sample values for the parameters. So the estimated regression equation is $\bar{y}+r\frac{s_y}{s_x}(x-\bar{x})$.

Notice also that the variance of the conditional distribution of *Y* given *X* = *x* is $\sigma_y{}^2(1-\rho^2)$, which is less than $\sigma_y{}^2$. This can be interpreted as saying that the uncertainty of *Y* is reduced by a factor of $\rho^2$ once the value of *X* is known. $\rho^2$ is therefore a measure of the value of the information that *X* has about *Y*. If $\rho^2$ = 0 (i.e., *X* and *Y* are independent under the bivariate normal distribution), then *X* has no information about *Y*. If $\rho^2$ = 1, then *X* has perfect information about *Y* and you can compute the value of *Y* once *X* = *x* is known, for then the conditional variance of *Y* is zero, so the value of *Y* equals the conditional mean. The sample (estimated) value of $\rho^2$ is $r^2$, a very important statistic called *R-square* in regression, which is interpreted as a measure of how well *X* explains *Y*.

## Section 4. Random Samples

**Sampling.** We often try to learn about a population by drawing repeatedly from it. A statistical model for this process can be set up by modeling each draw as a separate RV $X_i$, *i*=1, 2, …, *n*. Before we learn the result of the *i*th draw, the outcome is *potential*; upon learning the result, the outcome becomes *actual*. That is, the RV $X_i$ yields an outcome $x_i$, which should be distinguished from the RV $X_i$. The RV is not the outcome. The RV $X_i$ represents the set of potential outcomes and their probability distribution. The actual outcome $x_i$ is one of those potential outcomes. The outcome $x_i$ is often called a **realization** or **observation** or **instantiation** of the RV $X_i$. The collection of these RVs $(X_1, X_2, ..., X_n)$ is a **sample** in the potential sense. And the observed outcomes $(x_1, x_2, ..., x_n)$ are also a **sample**, but in the realized sense.

Once we have the sample, we compute statistics to estimate things. For example, we may compute the realized sample mean $\bar{x} = \dfrac{x_1 + \cdots + x_n}{n}$. But $\bar{x}$ can be considered a realization of a RV $\bar{X} = \dfrac{X_1 + \cdots + X_n}{n}$. The latter is a legitimate RV, the same as any other - $\bar{X}$ has a set of potential outcomes (all of the possible $\bar{x}$'s) and a probability function that assigns probabilities to the potential $\bar{x}$'s. Both the set of $\bar{x}$ outcomes and their probabilities are determined completely by the population of outcomes {( $x_1, x_2, ..., x_n$ )'s} for the random vector $(X_1, X_2, ..., X_n)$ and by the probability function $f(x_1, x_2, ..., x_n)$ that assigns probabilities to those outcomes.

What we mean by the **sampling distribution** of $\bar{X}$ is just the probability distribution of the random variable $\bar{X}$ - its outcome set and probability function. So, in spite of the difficulty that students have in grasping the concept of sampling distribution in lower-level courses, there is, *in principle*, nothing new for us to learn about the concept. However, the importance of sampling and, especially, the major importance of the statistics used in sampling – like the mean and variance – impel us to flesh out the development.

The following additional facts will be helpful in figuring out properties of sampling distributions.

- Suppose $(X_1, X_2, ..., X_n)$ are joint RVs with density function $f(x_1, x_2, ..., x_n)$. $(X_1, X_2, ..., X_n)$ are **mutually independent** if $f(x_1, x_2, ..., x_n) = f_{X_1}(x_1) f_{X_2}(x_2) \cdots f_{X_n}(x_n)$.
- A random sample is a sample that has the two special properties of independence and same distribution. A **random sample** is a set of independent random variables $(X_1, X_2, ..., X_n)$ with the same population (outcome set) and the same marginal probability distributions.[8] That is, a random sample (RS) is a set of **independent and identically distributed (i.i.d.)** random variables. The density function of a RS is $f(x_1, x_2, ..., x_n) = f_X(x_1) f_X(x_2) \cdots f_X(x_n) = \prod_{i=1}^{n} f_X(x_i)$, where $f_X(x)$ is the common marginal density function.

[8] "Random sample" may also refer to the realization $(x_1, x_2, ..., x_n)$, as well as to the RVs $(X_1, X_2, ..., X_n)$.

Ex. Suppose a coin is tossed 10 times. Let $X_i$ be 1 if a head is observed on toss *i*, and be 0 if a tail is observed, *i*=1, 2, …, 10. Since the coin is the same on each toss, it is reasonable to suppose that the probabilities of head and tail remain the same on all tosses. So the distribution remains the same. It is probably true as well that there is no connection between the tosses, so that the outcome of one toss does not affect the outcome of any other toss. Then the tosses are independent. If necessary, these suppositions can be tested. If confirmed or posited, then $(X_1, X_2, ..., X_{10})$ is a Random Sample. Note that if the coin were changed during the sequence of tosses – say by replacing it with a coin of different composition, perhaps with lead plating on one side – then the probabilities may change and the sequence would no longer be identically distributed. Note also that if each time the coin produces a head, the manner of flipping is changed – say to do a lazy one or two-turn flip in an attempt to produce a tail, then (assuming some success) it would be more likely to get a tail following a head than following a tail, so the tosses would be dependent to some degree.

Ex. Suppose an apartment is selected at random from a list of all Austin apartments. This means that each apartment has the same chance of selection as any other apartment. After selection of the first apartment, a second apartment is selected in the same manner from the remaining apartments on the list. This means that each remaining apartment has the same chance of selection as any other remaining apartment. Having already been selected, the first selected apartment is kept out of the second draw. This continues until 60 apartments have been selected. The 60 are all different apartments. Let $X_i$ be the monthly rent of apartment *i*, *i*=1, 2, …, 60. Are $(X_1, X_2, ..., X_{60})$ a Random Sample?

No. The probabilities change from one selection to another. For example, if there are (say) 200,000 apartments on the original list, then on the initial selection, the probability of choosing any apartment is 1/200,000. But after the selection of the first apartment, the probability of selection of any remaining apartment becomes 1/199,999 and the probability of a repeat selection of the first apartment becomes zero. Since the outcome of any selection changes the probabilities on subsequent selections, the selections are dependent.[9] However, it does not seem that selection affects the probabilities very much, except for those few apartments that have been selected. The difference between 1/200,000 and 1/199,999 or even 1/199,941 (for selection 60) is very small, although the difference between 1/199,941 and zero is relatively much larger, although still small in absolute terms.

If the selected apartments are returned to the list after selection and allowed to (possibly) be selected again, then the probabilities remain at 1/200,000 on each selection and $(X_1, X_2, ..., X_{60})$ a Random Sample. The latter type of sampling is called **sampling with replacement** and meets the requirements for a Random Sample. However, most real-world sampling is **sampling without replacement** for the obvious reason that once you have picked a sample item, you do not need to see it again. Yet, the difference between the two types is often slight, and the mathematics is easier for Random Samples (with replacement). In fact, it is common practice to treat sampling without replacement as though it is a Random Sample provided the math does not matter much. It does not matter much if the sample is small relative to the whole population from which it is drawn. As a rule of thumb, "small" in this context means less than 5% of the population. In the apartment illustration, selecting 60 apartments from 200,000 certainly qualifies as less than 5% of the population. If the sample size is not small in this sense, there are mathematical adjustments that can be made to compensate.

[9] It is less obvious that the selections are still identically distributed: Before any selections are made, the probability that a particular apartment will be chosen on selection number 23 (say) is 1/200,000.

## Section 5. Linear Combinations

*A hot tip!* Linear combinations and their properties are at the very heart of much of statistical estimation. Most estimators in regression are linear combinations or functions of linear combinations. Linear combinations are also at the heart of factor analysis, principal components analysis, discriminant analysis, logistic regression, canonical correlation, generalized linear models, etc. Developing an understanding of linear combinations will give you a head start on a lot of statistical methodology.

**Linear combinations.** A **linear combination** of RV's $X_1, X_2, \ldots, X_n$ is a univariate RV with the form $a_1X_1 + a_2X_2 + \ldots + a_nX_n$, where $a_1, a_2, \ldots, a_n$ are constants.

Ex. $X + Y = 1X + 1Y$

Ex. $X - Y = 1X + (-1)Y$

Ex. $X_1 = 1X_1 + 0X_2 + \ldots + 0X_n$

Ex. The sample mean $\overline{X} = \frac{1}{n}X_1 + \frac{1}{n}X_2 + \ldots + \frac{1}{n}X_n$ is a linear combination.

Ex. The sample variance $S^2 = \sum_{i=1}^{n}(X_i - \overline{X})^2/(n-1)$ is not itself a linear combination.[10] But it is a function of linear combinations, for

$$X_i - \overline{X} = \left(\frac{-1}{n}\right)X_1 + \ldots + \left(\frac{-1}{n}\right)X_{i-1} + \left(1 - \frac{1}{n}\right)X_i + \left(\frac{-1}{n}\right)X_{i+1} + \ldots + \left(\frac{-1}{n}\right)X_n$$ is a linear combination.

The importance of linear combinations in statistics cannot be overstated. Many statistical estimators are linear combinations of RVs. Most of regression and multivariate statistics would not exist without linear combinations. Here are some fundamental useful properties of linear combinations. Make special note of these.

- Mean of a linear combination:
  $E(a_1X_1 + a_2X_2 + \ldots + a_nX_n) = a_1E(X_1) + a_2E(X_2) + \ldots + a_nE(X_n)$. In words, the mean of a linear combination is the linear combination of the means. This holds whether or not the *X*'s are independent and/or identically distributed.

  Ex. The mean of the sample mean of a RS: $E(\overline{X}) = \frac{1}{n}E(X_1) + \frac{1}{n}E(X_2) + \ldots + \frac{1}{n}E(X_n) = \frac{1}{n}\mu + \frac{1}{n}\mu + \ldots + \frac{1}{n}\mu = \mu$, where $\mu$ is the common mean.

- Variance of a linear combination:

  $$Var(a_1X_1 + a_2X_2 + \ldots + a_nX_n) = a_1^2Var(X_1) + a_2^2Var(X_2) + \ldots + a_n^2Var(X_n) + \sum_{i \neq j}\sum a_ia_jCov(X_i.X_j)$$

  Ex. The variance of the sum of a RS is $Var(X_1 + X_2 + \ldots + X_n) = Var(X_1) + Var(X_2) + \ldots + Var(X_n) = n\sigma^2$, where $\sigma^2$ is the common variance. The Cov terms are zeroes because of independence. In words, the variance of a sum of RV's is the sum of the variances when the RVs are independent.

[10] It is a quadratic form.

Ex. The variance of the sample mean of a RS is $Var\left(\frac{X_1 + X_2 + ... + X_n}{n}\right) = \frac{\sigma^2}{n}$.

Ex: Suppose that an inheritance is invested in equal portions among three stocks, with prices $X_1, X_2, X_3$ that have the same variance $\sigma^2$ and equal pairwise correlation = 0.5. Then the price volatility of the combined portfolio may be assessed by $Var\left(\frac{X_1 + X_2 + X_3}{3}\right)$ = $(1/3)^2 Var(X_1) + (1/3)^2 Var(X_2) + (1/3)^2 Var(X_n) + \sum\sum_{i \neq j} (1/3)(1/3)Cov(X_i.X_j)$ = $(3/9)\sigma^2 + \sum\sum_{i \neq j} (1/9)0.5\sigma\sigma$ = $(3/9)\sigma^2 + (3/9)\sigma^2 = (2/3)\sigma^2$. So the price volatility of the diversified portfolio is reduced compared with a portfolio invested entirely in one of the stocks ($\sigma^2$) but not as much as if the three stock prices had been independent ($(1/3)\sigma^2$). This is a general property of diversification under more realistic conditions of unequal partitioning, unequal variances, and unequal correlations.

- If $(X_1, X_2, ..., X_n)$ is a RS in which each $X_i$ has a normal distribution with mean $\mu$ and variance $\sigma^2$, then $a_1 X_1 + a_2 X_2 + ... + a_n X_n$ has a normal distribution with mean $(a_1 + ... + a_n)\mu$ and variance $(a_1^2 + ... + a_n^2)\sigma^2$.

  Ex. With the same assumption, it follows that the sample mean $\overline{X}$ has a normal distribution with mean $\mu$ and variance $\sigma^2 / n$.

For figuring out the relationship between two linear combinations, the following property is key.

- $Cov(aX + bY, cU + dV) = acCov(X,U) + adCov(X,V) + bcCov(Y,U) + bdCov(Y,V)$

Make special note of the preceding bulleted item. It can be used to derive a large number of rules for special cases. For example, all of the following are consequences:

- $Cov(aX + bY, cX + dY) = acVar(X) + (ad + bc)Cov(X,Y) + bdVar(Y)$
- $Var(aX + bY) = a^2 Var(X) + 2abCov(X,Y) + b^2 Var(Y)$
- If *X* and *Y* are independent, then $Var(aX + bY) = a^2 Var(X) + b^2 Var(Y)$
- In particular, if *X* and *Y* are independent, then $Var(X + Y) = Var(X) + Var(Y)$ and $Var(X - Y) = Var(X) + Var(Y)$; and if *X* and *Y* are dependent, then $Var(X + Y) = Var(X) + 2Cov(X,Y) + Var(Y)$
- $Corr(a + bX, c + dY) = sign(bd)Corr(X,Y)$. That is, correlation is unaffected by a linear transformation of the variables, as long as the coefficients of *X* and *Y* have the same sign.

## Section 6. Estimation.

The purpose of calculating sample statistics is usually to estimate population parameters. For example, the sample mean estimates the population mean. The sample variance estimates the population variance. But are these good estimates? Are they sometimes not so good? What motivates them?

A large number of methods have been proposed for producing good estimators. I will discuss three methods: ordinary least squares (OLS), maximum likelihood, and Bayes estimation. Each method produces estimators that are optimal according to an appealing criterion.

Estimation begins with a **statistical model**, which is really a set of **specifications** of what is believed to hold about the random variables under observation or experimentation. The model specifies some things precisely (e.g., that the RVs all have variances equal to 1) and other things imprecisely (e.g., that the means of the RVs must be positive numbers). The specifications of the linear regression model are especially important. Values that play a role in the model but that are not completely specified are called **parameters**. For example, in simple regression, the linear specification is that the conditional mean of *Y* given *X* = *x* is $\alpha + \beta x$. This specifies the general form but does not specify definite values for the intercept $\alpha$ and slope $\beta$, which are parameters. Estimation is about guessing the specific values of the parameters.

OLS, maximum likelihood, and Bayes are three general approaches to estimating model parameters. The following is a sound-bite synopsis of each method:

- OLS tries to pick an estimate that is closest to the data.
- Maximum likelihood tries to pick the most likely estimate.
- Bayes tries to adjust your prior opinions in the light of the data.

**Note on notation.** I intend to use the term **estimate** to refer to the actual value used to estimate a parameter. I will use the term **estimator** to refer to the potential value of the estimate, prior to completing the observations. This distinction should recall my earlier distinction between random variable *X* and realization *x*. In fact, an estimator is a random variable and an estimate is a realization of that random variable.

Ex: Take a random sample $X_1, X_2, \ldots, X_n$. The sample mean $\bar{X}$ is an *estimator* of the population mean $\mu$. If you observe the values $x_1, x_2, \ldots, x_n$ of the random variables and calculate the observed value of $\bar{X}$ to be $\bar{x}$ = 6, then 6 is an *estimate* of $\mu$.

Every estimator must be a statistic. A **statistic** is a random variable, the calculation of the value of which does not depend upon the value of any unknown parameter. However, the distribution of a statistic may depend upon the value of an unknown parameter. The realized value of a statistic is also called a statistic, in a possible abuse of terminology.

Ex: Take a random sample $X_1, X_2, \ldots, X_n$ from a normal distribution with *unknown* mean $\mu$ and *known* variance $\sigma^2$. Then the sample mean $\bar{X}$ is a statistic. $\bar{x}$ is a statistic. The sample variance $S^2$ is a statistic. $\dfrac{(n-1)S^2}{\sigma^2}$ is a statistic. $\dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ is *not* a statistic. Although the latter is not a statistic, it is still a random variable. Note that although the *calculation* of the latter depends upon an unknown parameter, the *distribution* of the latter does not (it is standard normal); such a random variable is called a pivot – useful in making confidence intervals and testing hypotheses.

Since we actually want to calculate our statistics, we must be able to complete the calculation, so that calculation cannot depend upon the value of any unknown parameter. That is why every estimator must be a statistic.

**1. OLS.**

Least squares is not just for regression. Least squares is a general method for producing estimators. In OLS, given the general form of the model, you pick values for the model parameters that are closest to the data in the sense of least squares.

Here is some more detail on the general case for Least Squares, followed by two examples. General Case: Suppose that we have realizations $x_1, x_2, ..., x_n$ of the RVs $X_1, X_2, ..., X_n$ and that the model specifies $g(\theta_1, \theta_2, ..., \theta_p)$ as an expected form for the data. That is, each $x_i$ "should" equal $g(\theta_1, \theta_2, ..., \theta_p)$, according to the model. The specification *g* is a function of $\theta_1, \theta_2, ..., \theta_p$, which are all of the parameters to which the model does not give specific values. *g* may also be a function of given constants, including data (realizations of the RVs). The **principle of least squares** says to choose specific values for the parameters to minimize the error sum of squares (ESS, also called total squared error) $\sum_{i=1}^{n} \{x_i - g(\theta_1, \theta_2, ..., \theta_p)\}^2$. Since $x_i$ "should" equal $g(\theta_1, \theta_2, ..., \theta_p)$, then $x_i - g(\theta_1, \theta_2, ..., \theta_p)$ is the deviation, or error, between the actual value $x_i$ and the model value $g(\theta_1, \theta_2, ..., \theta_p)$. The ESS criterion defines what it means to be "close to the data" in the OLS method. Usually – but not always – the error sum of squares lends itself to minimization by differentiation. The values $\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_p$ that minimize the error sum of squares are called **(ordinary) least squares estimates (OLS)**. The corresponding random variables whose realizations are the OLS estimates are the **(ordinary) least squares estimators** (**OLS**).

Ex: OLS is not just for regression. As an example, suppose we want to estimate the mean $\mu$ of a certain distribution. Suppose we observe realizations $x_1, x_2, ..., x_n$ of the RVs $X_1, X_2, ..., X_n$, for which the model specifications are that the RVs are identically distributed with the certain distribution as their common distribution, and the unknown mean $\mu$ of that distribution could be any value. Then each observed $x_i$ "*should be*" the common mean $\mu$. That is, $g(\theta) = g(\mu) = \mu$. The only reason that $x_i$ is not $\mu$ is random error. Then the error sum of squares is $\sum_{i=1}^{n} \{x_i - \mu\}^2$. Minimization of the ESS over all possible values of $\mu$ yields $\hat{\mu} = \bar{x}$ as the least-squares estimate. To minimize ESS, differentiate it and set the derivate equal to 0 to find critical points: $\frac{d}{d\mu} \sum_{i=1}^{n} \{x_i - \mu\}^2 = \sum_{i=1}^{n} \frac{d}{d\mu} \{x_i - \mu\}^2 = \sum_{i=1}^{n} (-2)\{x_i - \mu\}$ set = 0 yields $-2\sum_{i=1}^{n} x_i + 2\sum_{i=1}^{n} \mu = 0$. Hence $\sum_{i=1}^{n} x_i = n\mu$. So $\hat{\mu} = \bar{x}$ is a critical point, and the second derivative is positive, so the solution is a minimizer. The corresponding RV whose realizations are the least-squares estimate is the least-squares estimator $\bar{X}$.

Ex: In the simple linear regression model, we observe realizations $y_1, y_2, ..., y_n$ of the RVs $Y_1, Y_2, ..., Y_n$ at given *x*-values $x_1, x_2, ..., x_n$ (in ordinary regression, the *x*'s are not RVs). The simple linear regression model specifications say that each realized *y* "should be" $g(\alpha, \beta) = \alpha + \beta x$ and would be, but for random error. The values of *α* and *β* are not specified, but the value of *x* is. If we observe $y_i$, the error is $y_i - (\alpha + \beta x_i)$. So OLS chooses *α* and *β* to minimize ESS = $\sum_{i=1}^{n} \{y_i - (\alpha + \beta x_i)\}^2$. In regression, the ESS is the sum of squared residuals.

## 2. Maximum likelihood.

The statistical model usually does not completely specify the probability distribution of the RVs $X_1, X_2, ..., X_n$ to be observed. Their density function $f(x_1, x_2, ..., x_n)$ depends upon parameters $\theta_1, \theta_2, ..., \theta_p$. To make this dependence explicit, we can write $f(x_1, x_2, ..., x_n; \theta_1, \theta_2, ..., \theta_p)$ for the joint density. Once we have observed the realizations $x_1, x_2, ..., x_n$ of the RVs $X_1, X_2, ..., X_n$, the realizations are known. Then we can substitute the realizations into the density and view the joint density $f(x_1, x_2, ..., x_n; \theta_1, \theta_2, ..., \theta_p)$ as a function of $\theta_1, \theta_2, ..., \theta_p$ instead of a function of $x_1, x_2, ..., x_n$. This gives us the **likelihood function** $L(\theta_1, \theta_2, ..., \theta_p; x_1, x_2, ..., x_n) = f(x_1, x_2, ..., x_n; \theta_1, \theta_2, ..., \theta_p)$. That is, *the likelihood function is just the joint density viewed as a function of the parameters rather than as a function of the data.* The justification for this view is that once the RVs have been observed, the parameters are the only unknowns in the density function.

In the discrete case, the joint density is the probability of observing $x_1, x_2, ..., x_n$. If we view the joint density as a function of the parameters, then the probability of observing $x_1, x_2, ..., x_n$ varies with the parameter values. So a natural question is, For which specific set of parameter values is the probability of observing the actual data $x_1, x_2, ..., x_n$ at a maximum? That particular set of parameter values is the set that is most likely to have produced the data $x_1, x_2, ..., x_n$ that we observed. That set of parameter values $\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_p$ is called the **maximum likelihood estimate** (**mle**). It is important to note that the mle is a parameter value, but the maximizing value may vary, depending on the observed realizations $x_1, x_2, ..., x_n$. Thus, the mle is a function of the observed data $x_1, x_2, ..., x_n$. The corresponding random variable whose realizations are the maximum likelihood estimate is the **maximum likelihood estimator** (**MLE**).

In the continuous case, the joint density is not the *probability* of observing $x_1, x_2, ..., x_n$. So we do not have the nice interpretation of maximum likelihood as choosing the parameter values under which the observed data were most likely to have occurred. But what we have in the continuous case is pretty close to this. In the continuous case, the p.d.f. is the concentration of probability at $x_1, x_2, ..., x_n$. Given observed $x_1, x_2, ..., x_n$, we can choose the set of parameter values $\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_p$ that maximizes the concentration of probability at the observed $x_1, x_2, ..., x_n$. We speak of maximizing the "*likelihood*" (referring to the *likelihood function* and mean maximizing the concentration of probability) instead of maximizing the probability. We can often find the mle by differentiating the likelihood function with respect to the parameters.

Ex: Suppose we observe realizations $x_1, x_2, ..., x_n$ of the RVs $X_1, X_2, ..., X_n$ and the model is the random sample model. Suppose that the model further specifies that the RVs $X_1, X_2, ..., X_n$ are Bernoulli, with P($X$ = 1) = $\pi$ and P($X$ = 0) = 1 - $\pi$. Then the pdf of each can be written compactly as $f(x_i;\pi) = \pi^{x_i}(1-\pi)^{1-x_i}$ for $x_i$ = 0 or 1 (Verify this.) Our job is to estimate $\pi$ by maximum likelihood. The likelihood function is $L(\pi; x_1, x_2, ..., x_n)$ = $\prod_{i=1}^n \pi^{x_i}(1-\pi)^{1-x_i}$ = $\pi^{\sum x_i}(1-\pi)^{n-\sum x_i}$. To maximize, calculate $\frac{d}{d\pi}L(\pi; x_1, x_2, ..., x_n)$ = $\sum x_i \pi^{\sum x_i - 1}(1-\pi)^{n-\sum x_i} + \pi^{\sum x_i}(n - \sum x_i)(1-\pi)^{n-\sum x_i - 1}(-1)$ and set = 0. Then $\sum x_i(1-\pi) + \pi(n - \sum x_i)(-1)$ = 0. Solving, we get $\pi = \frac{\sum x_i}{n}$, which can be verified to be a point of maximum in [0,1]. The maximum likelihood estimate is $\hat{\pi} = \frac{\sum x_i}{n}$, i.e., the sample proportion of 1's. Since this calculation works whatever the values of $x_1, x_2, ..., x_n$ may be, then the corresponding RV, $\hat{\pi} = \frac{\sum X_i}{n}$, is the maximum likelihood estimator.

Ex: Suppose we want to estimate the mean $\mu$ and variance $\sigma^2$ of a certain distribution by maximum likelihood. Suppose we observe realizations $x_1, x_2, ..., x_n$ of the RVs $X_1, X_2, ..., X_n$, which are specified to be independent with the certain distribution as their common distribution (the random sample model). Suppose that the model further specifies that the RVs $X_1, X_2, ..., X_n$ are normal with unknown mean $\mu$ and unknown variance $\sigma^2 > 0$. Then the joint density is $f(x_1, x_2, ..., x_n; \mu, \sigma^2)$ = $\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x_1-\mu)^2}{2\sigma^2}} \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x_2-\mu)^2}{2\sigma^2}} \cdots \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x_n-\mu)^2}{2\sigma^2}}$ = $\sigma^{-n}(2\pi)^{-n/2}e^{-\sum_{i=1}^n \frac{(x_i-\mu)^2}{2\sigma^2}}$. Thus the likelihood function is $L(\mu, \sigma^2; x_1, x_2, ..., x_n) = \sigma^{-n}(2\pi)^{-n/2}e^{-\sum_{i=1}^n \frac{(x_i-\mu)^2}{2\sigma^2}}$. This can be maximized by calculus (remember that $L$ is a function of $\mu$ and $\sigma^2$ – the data values $x_1, x_2, ..., x_n$ are treated as constants), yielding the maximum likelihood estimates $\hat{\mu} = \bar{x}$ and $\hat{\sigma}^2 = \sum_{i=1}^n (x_i - \bar{x})^2 / n$. The corresponding maximum likelihood estimators are $\bar{X}$ and $\sum_{i=1}^n (X_i - \bar{X})^2 / n$ – the sample mean and sample variance (but dividing by $n$ instead of $n$ – 1).

Ex: Suppose we observe realizations $x_1, x_2, ..., x_n$ of the RVs $X_1, X_2, ..., X_n$ and the model is the random sample (i.i.d.) model. Suppose that the model further specifies that the RVs $X_1, X_2, ..., X_n$ have common pdf $f(x_i;\theta) = e^{-(x_i-\theta)}$ for $x_i \geq \theta$, which we can write for all $x_i$ as $f(x_i;\theta) = e^{-(x_i-\theta)}I_{[\theta,\infty)}(x_i)$, where $I_{[\theta,\infty)}(x_i)$ = 1 if $x_i \geq \theta$ and = 0 if $x_i < \theta$. So the likelihood function is $L(\theta; x_1, x_2, ..., x_n)$ = $\prod_{i=1}^n e^{-(x_i-\theta)}I_{[\theta,\infty)}(x_i) = e^{-\sum (x_i-\theta)}\prod_{i=1}^n I_{[\theta,\infty)}(x_i)$ =

$e^{-\sum (x_i-\theta)} I_{[\theta,\infty)}(\min(x_1,\ldots,x_n))$ (Verify this.) Unfortunately, $L(\theta; x_1, x_2,\ldots,x_n)$ is not differentiable at all $\theta$. It has a discontinuity at $\theta = \min(x_1,\ldots,x_n)$ (Verify this.) However, it is easy to see that $L(\theta; x_1, x_2,\ldots,x_n) = 0$ for all $\theta > \min(x_1,\ldots,x_n)$ and increases with $\theta$ whenever $\theta < \min(x_1,\ldots,x_n)$. Thus, $L(\theta; x_1, x_2,\ldots,x_n)$ is maximized at $\theta = \min(x_1,\ldots,x_n)$, which is the mle.

## 3. Bayes estimation.

Bayes estimation starts by inquiring what your opinions are of the likely values for the model parameters $\theta_1,\theta_2,\ldots,\theta_p$. You should be able to express those opinions[11] in the form of a density function $f(\theta_1,\theta_2,\ldots,\theta_p)$, called the **prior distribution** of the parameters. (Notice that the prior distribution is a probability distribution for the parameters – it is *not* a probability distribution for the data.) Then you perform the experiment or observation. That is, you observe realizations $x_1, x_2,\ldots,x_n$ of RVs $X_1, X_2,\ldots,X_n$ whose distribution depends on $\theta_1,\theta_2,\ldots,\theta_p$. You then revise your prior opinions about the parameters in light of the new data. This gives you a **posterior distribution** $f(\theta_1,\theta_2,\ldots,\theta_p \mid X_1 = x_1, X_2 = x_2,\ldots,X_n = x_n)$ of the parameters (not of the data). Some appropriate characteristic of the posterior distribution is then chosen as the **Bayes estimate** of $\theta_1,\theta_2,\ldots,\theta_p$. Quite often the mean of the posterior distribution is used. Note that the posterior mean is a function of the observed data $x_1, x_2,\ldots,x_n$. The corresponding random variable whose realizations are the Bayes estimate is the **Bayes estimator** of $\theta_1,\theta_2,\ldots,\theta_p$.

The key to Bayes estimation is how to combine the prior opinion with the data to yield the posterior opinion. That is done by an application of Bayes Theorem: The posterior density of the parameters is obtained by $f(\theta_1,\theta_2,\ldots,\theta_p \mid X_1 = x_1, X_2 = x_2,\ldots,X_n = x_n)$ =

$$\frac{f(\theta_1,\theta_2,\ldots,\theta_p, X_1 = x_1, X_2 = x_2,\ldots,X_n = x_n)}{f(X_1 = x_1, X_2 = x_2,\ldots,X_n = x_n)} =$$

$$\frac{f(X_1 = x_1, X_2 = x_2,\ldots,X_n = x_n \mid \theta_1,\theta_2,\ldots,\theta_p)\, f(\theta_1,\theta_2,\ldots,\theta_p)}{f(X_1 = x_1, X_2 = x_2,\ldots,X_n = x_n)} =$$

$$\frac{L(\theta_1,\theta_2,\ldots,\theta_p; x_1, x_2,\ldots,x_n)\, f(\theta_1,\theta_2,\ldots,\theta_p)}{f(X_1 = x_1, X_2 = x_2,\ldots,X_n = x_n)}$$

. The numerator of the latter is the product of the likelihood function (which is provided by the statistical model) and the prior density (which is provided by the investigator's opinions) – so both parts in the numerator are available. The denominator is the marginal density of the data, which does not depend on $\theta_1,\theta_2,\ldots,\theta_p$, and can be obtained from the numerator, which is the joint density of the data $x_1, x_2,\ldots,x_n$ and the parameters $\theta_1,\theta_2,\ldots,\theta_p$. In the continuous case, the denominator (the marginal density of the data) is obtained by integration:

[11] These opinions could be mere conjecture, but should represent best available information, including past data, prior to the experiment or observation that you now contemplate.

$f(X_1 = x_1, X_2 = x_2, ..., X_n = x_n) = \int \cdots \int f(\theta_1, \theta_2, ..., \theta_p, X_1 = x_1, X_2 = x_2, ..., X_n = x_n)\, d\theta_1 \cdots d\theta_p =$ $\int \cdots \int L(\theta_1, \theta_2, ..., \theta_p; x_1, x_2, ..., x_n)\, f(\theta_1, \theta_2, ..., \theta_p)\, d\theta_1 \cdots d\theta_p$ . That is, the marginal density of the data is obtained by "integrating out" the parameters in the joint density of both data and parameters. In the discrete case, integrals are replaced by summations. Thus, the prior opinion and the likelihood function provide everything necessary to calculate Bayes estimators.

The calculation of Bayes estimators by hand is tedious and hard except in special cases. For that reason, Bayesian statistics was a minor backwater loudly championed by a few vocal advocates, but rather impractical for most serious work. Around 1990, however, two developments converged that completely reversed the prospects for Bayesian statistics. One was the development of theory that showed that the Bayesian program could be implemented in the hard cases by relatively easy simulations. The other was the arrival of sufficiently powerful computers that could do the necessary large-scale simulations in reasonable time. Since then, Bayesian statistics has flourished.

I will illustrate Bayesian estimation with a few simple (non-simulation) examples.

> Ex: Suppose that we want to estimate the mean $\mu$ of a certain distribution. Suppose that we have RVs $X_1, X_2, ..., X_n$ , and the model is the random sample model. Suppose that the model further specifies that the RVs $X_1, X_2, ..., X_n$ are normal with unknown mean $\mu$ and known variance $\sigma^2 = 1$. Suppose that the prevailing opinion is that the most likely value for the unknown mean $\mu$ is 100, but that values close to 100 are almost as probable, with the likelihood tapering off symmetrically in either direction from 100. We may, in fact, express this opinion about $\mu$ in the form of a prior distribution that is normal, with a mean of 100 and a variance of 2. We then observe data values $x_1, x_2, ..., x_n$ and revise our prior opinion. What is our posterior opinion? To calculate the posterior distribution of $\mu$ given $x_1, x_2, ..., x_n$, we need the prior density and the likelihood. The prior density is $f(\mu) = \frac{1}{\sqrt{2}\sqrt{2\pi}} e^{-\frac{(\mu-100)^2}{4}}$ . The likelihood function is
>
> $$L(\mu; x_1, x_2, ..., x_n) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x_1-\mu)^2}{2}} \frac{1}{\sqrt{2\pi}} e^{-\frac{(x_2-\mu)^2}{2}} \cdots \frac{1}{\sqrt{2\pi}} e^{-\frac{(x_n-\mu)^2}{2}} = (2\pi)^{-n/2} e^{-\sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2}} .$$
>
> With some tedious algebra, it can be shown that the posterior density $f(\mu \mid X_1 = x_1, X_2 = x_2, ..., X_n = x_n)$ is normal with mean $\frac{100/n + 2\bar{x}}{2 + 1/n}$ and variance $\frac{2/n}{2 + 1/n}$.
>
> The Bayes estimate of $\mu$ is the posterior mean $\frac{100/n + 2\bar{x}}{2 + 1/n} = 100\frac{1/n}{2 + 1/n} + \bar{x}\frac{2}{2 + 1/n}$, which is a weighted average of the prior opinion (100) about $\mu$ with the data's opinion ( $\bar{x}$ ) about $\mu$ . The weights are proportions obtained from the prior variance (2) and the data variance (1/n). Thus the Bayes estimate is partway between the prior mean and the data mean. By looking at the weights, you can see that as the sample size increases, the weight on the prior opinion (100) reduces toward zero and the weight on the data opinion ( $\bar{x}$ ) increases toward 1. So as the sample size increases, the Bayes estimate gets closer to the sample mean. So in the limit, the effect of the prior opinion is overwhelmed by the weight of the data. This is a general property of Bayes estimates.

Ex: Suppose that you have received a tip that your favorite casino has a defective slot machine. When you bet \$1, the casino's good slot machines pay out \$5 twenty percent of the time, and pay out nothing the rest of the time. But the defective machine pays out \$5 forty percent of the time when you bet \$1, and pays out nothing the rest of the time. You think there is a 50-50 chance that the slot machine in front of you is the defective machine. You bet \$1 and win \$5. How should you revise your opinion? Can you now be confident that your slot machine is your ticket to Rich City? In this case, the parameter is the status of the slot machine in front of you. Let $\theta$ = 1 if the machine is defective, = 0 otherwise. Then your prior distribution is $P(\theta=1) = 0.5$, $P(\theta=0) =$ 0.5. Let *X* = the payout when you bet. The likelihood function can be expressed compactly as $L(\theta;x) = (0.2+0.2\theta)^{x/5}(0.8-0.2\theta)^{1-x/5}$. That is, $L(\theta=0;x) = (0.2)^{x/5}(0.8)^{1-x/5}$, which is 0.8 if your payoff is *x* = 0 and is 0.2 if your payoff is *x* = 5; and $L(\theta=1;x) = (0.4)^{x/5}(0.6)^{1-x/5}$, which is 0.6 if your payoff is *x* = 0 and is 0.4 if your payoff is *x* = 5. After you have bet \$1 and obtained your payout of *x*, then you should revise your prior opinion to become the posterior

$$f(\theta\,|\,x) = \frac{L(\theta;x)f_\theta(\theta)}{f_X(x)} = \frac{(0.2+0.2\theta)^{x/5}(0.8-0.2\theta)^{1-x/5}0.5}{\sum_{\theta=0}^{1}(0.2+0.2\theta)^{x/5}(0.8-0.2\theta)^{1-x/5}0.5} =$$

$$\frac{(0.2+0.2\theta)^{x/5}(0.8-0.2\theta)^{1-x/5}}{(0.2)^{x/5}(0.8)^{1-x/5}+(0.4)^{x/5}(0.6)^{1-x/5}} = \frac{0.8-0.2\theta}{1.4}$$ if *x* = 0 and = $\frac{0.2+0.2\theta}{0.6}$ if *x* = 5. So if you get \$5 from your toss, the posterior probability that you are playing the defective slot is $\frac{0.2+0.2*1}{0.6}$ = 2/3, and the posterior probability that you are playing an ordinary slot is 1/3. The larger of these posterior probabilities is 2/3. So if you use the posterior mode as your Bayesian estimate of $\theta$, you would estimate that you are playing the defective slot. That is, if you get a payoff of \$5, then it is more likely that you played the defective slot than the ordinary slot. On the other hand, if you get \$0 from your toss, the posterior probability that you are playing the defective slot is $\frac{0.8-0.2*1}{1.4}$ = 3/7, and the posterior probability that you are not is 4/7. So if you use the posterior mode as your Bayesian estimate of $\theta$ and you get \$0, you would estimate that you are playing an ordinary slot. That is, if you get a payoff of \$0, then it is more likely that you played the ordinary slot than the defective slot. In neither case should you feel very confident. If you continue to test the machine by playing it repeatedly, you will collect more observations and the analysis will become more complicated (see next).

Ex: (Continuation) Suppose that your prior opinion is 0.5 that your slot machine is defective and 0.5 that it is an ordinary machine. Suppose that you play the slot machine *n* times. The posterior distribution is $f(\theta\,|\,x_1,\ldots,x_n) = \frac{(0.2+0.2\theta)^{\sum x/5}(0.8-0.2\theta)^{n-\sum x/5}0.5}{\sum_{\theta=0}^{1}(0.2+0.2\theta)^{\sum x/5}(0.8-0.2\theta)^{n-\sum x/5}0.5}$ =

$\frac{(0.2+0.2\theta)^{\sum x/5}(0.8-0.2\theta)^{n-\sum x/5}}{\sum_{\theta=0}^{1}(0.2+0.2\theta)^{\sum x/5}(0.8-0.2\theta)^{n-\sum x/5}}$. Let us say that you have played 10 times and have been paid $\sum_{i=1}^{10} x_i$ = \$20 (4 wins out of 10). Then the posterior probability

$$f(\theta=1\,|\,x_1,\ldots,x_n) = \frac{(0.2+0.2*1)^4(0.8-0.2*1)^6}{(0.2+0.2*0)^4(0.8-0.2*0)^6+(0.2+0.2*1)^4(0.8-0.2*1)^6} =$$

$\dfrac{(0.4)^4(0.6)^6}{(0.2)^4(0.8)^6+(0.4)^4(0.6)^6}$ = 0.7401, and $f(\theta = 0 \mid x_1,\ldots,x_n)$ = 0.2599. So the posterior mode is the defective slot, but the evidence is not overwhelming.

**Some properties of estimators.**

Recall that an estimator is a RV. An estimator $\hat{\theta}$ is an **unbiased** estimator of $\theta$ if the mean of the estimator is $\theta$, i.e., $E(\hat{\theta}) = \theta$ for all possible values of $\theta$. The **bias** is $E(\hat{\theta}) - \theta$. An unbiased estimator equals its target parameter, on average. So unbiasedness is, in general, a good property for an estimator.

The **mean-squared error** (**MSE**) of an estimator is $MSE(\hat{\theta}) = E(\{\hat{\theta} - \theta\}^2)$ and is equal to $Var(\hat{\theta}) + [bias(\hat{\theta})]^2$. MSE is a measure of distance between an estimator and its target parameter.

An estimator $\hat{\theta}_1$ is more **efficient** than another estimator $\hat{\theta}_2$ if $MSE(\hat{\theta}_1) < MSE(\hat{\theta}_2)$ for all possible $\theta$. A more efficient estimator is closer to its target, on average. If an estimator is unbiased, then the MSE equals the variance. If both estimators are unbiased, then efficiency means that one has smaller variance than the other.

A **consistent** estimator gets closer to its target parameter as the sample size increases. There are different ways in which an estimator can get close to a parameter. One way is for its MSE to converge to zero. Another way is for the values of the estimator to converge to the target. Another is for the probabilities of the estimator to concentrate around the target. The latter sense is called **consistency in probability**: If the sequence of probabilities $P(\theta - \varepsilon < \hat{\theta} < \theta + \varepsilon)$ converges to one as the sample size $n \to \infty$ and if this happens for every $\varepsilon > 0$, no matter how small, then $\hat{\theta}$ is consistent in probability for $\theta$. Then we write $p\lim(\hat{\theta}) = \theta$.

**The law of large numbers**. Suppose that the random sample model applies, so that $X_1, X_2, \ldots, X_n$ is a RS and the common mean and standard deviation are $\mu$ and $\sigma^2$. Then the sample mean is consistent in probability for $\mu$ [12] and we write $p\lim(\overline{X}) = \mu$.

[12] Technical note: A stronger version says that the actual values of the sample mean converge to $\mu$, not just that the probabilities of $\overline{X}$ concentrate around $\mu$, except for a very few actual values having probability zero. This is called **almost sure consistency**. There are relationships among these forms of consistency: almost sure consistency => consistency in probability, and consistency in MSE => consistency in probability. But the relationship between almost sure consistency and consistency in MSE is more complex.

**Asymptotic distribution**. Suppose $\hat{\theta}_n$ is an estimator based on the first *n* RVs $X_1, X_2,..., X_n$ in a sequence and *Z* is a RV with a certain probability distribution. If the sequence of probabilities $P(a < \hat{\theta}_n < b)$ converges to $P(a < Z < b)$ for all *a* and *b*, then $\hat{\theta}_n$ converges in distribution to *Z*. That is, the probabilities for $\hat{\theta}_n$ get close to the corresponding probabilities for *Z*.[13]

The most useful application of asymptotic distribution is the **Central Limit Theorem**. There are several versions. Here is one: Suppose the random sample model applies, so that $X_1, X_2,..., X_n$ is a RS and the common mean and standard deviation are $\mu$ and $\sigma^2$, respectively. Then $\frac{\overline{X} - \mu}{\sigma / \sqrt{n}}$ converges in distribution to a standard normal RV. What this means is that the probabilities $P(a < \frac{\overline{X} - \mu}{\sigma / \sqrt{n}} < b)$ get close to $P(a < Z < b)$ as *n* gets big, for any given *a* and *b*, where *Z* has the standard normal density $\frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$. The practical application is that we can replace the complicated $P(a < \frac{\overline{X} - \mu}{\sigma / \sqrt{n}} < b)$ by the simple $P(a < Z < b)$ for a sufficiently large sample size and not be too far off. Loosely, and somewhat imprecisely, it is common to say that $\overline{X}$ has an approximate normal distribution with mean $\mu$ and variance $\sigma^2 / n$ if *n* is sufficiently large. It is surprising how small "sufficiently large" can be – in most applications, *n* > 30 is sufficient.

> Ex: Suppose $T_n$ is a RV that has Student's *t* distribution with *n* degrees of freedom and *Z* is a standard normal distribution. Then $P(a < T_n < b) \to P(a < Z < b)$ for all *a*, *b* as $n \to \infty$. That is, probabilities for Student's *t* converge to standard normal probabilities. This is why we can dispense with Student's *t* distribution and use the simpler normal distribution in computing *p*-values and in constructing confidence intervals if *n* is sufficiently large.
>
> Ex: Suppose $X_n$ has a binomial distribution with parameters *p* and *n*. Then $P\left( a < \frac{X_n - np}{\sqrt{np(1-p)}} < b \right) \to P(a < Z < b)$ for all *a*, *b* as $n \to \infty$. That is, probabilities for the standardized binomial converge to standard normal probabilities. Loosely, and somewhat imprecisely, it is common to say that $X_n$ has an approximate normal distribution with mean *np* and variance *np(1-p)* if *n* is sufficiently large. For the normal approximation to be reasonably good, $\min\{np, n(1-p)\}$ should be at least 5.

[13] Although the concept of asymptotic distribution is stated here for estimators, it is also used in exactly the same way for random variables that are not necessarily used as estimators.

Your statistical chores are not finished when you have calculated the value of an estimate of a parameter. You should also summarize the uncertainty of your estimate. There are two common ways of doing this:

- by reporting the standard error of your estimate, and/or
- by reporting a confidence interval.

Recall once again that an estimator is a RV. Your estimate is a realization of that estimator – one of the possible outcomes of that RV. The **standard error** of your estimate is just the standard deviation of your estimator. This is a parameter value, which is almost never known in practice. Therefore, common practice is to report an estimate of the standard error parameter. This estimate is the (estimated) standard error. The standard error can be interpreted as an estimate of the "average" amount by which your estimator can be expected to miss the parameter value that it is estimating.

> Ex: Suppose that we have RVs $X_1, X_2, ..., X_n$, and the model is the random sample model. Suppose that the model further does not specify the probability distribution of RVs $X_1, X_2, ..., X_n$ but does specify that the mean $\mu$ could be any number and the variance $\sigma^2$ could be any positive number. Suppose that we estimate $\mu$ by the sample mean $\bar{x}$. The standard error of $\bar{x}$ is the standard deviation of the RV $\bar{X}$, which is $\sigma/\sqrt{n}$. The value of $\sigma$ is not known but can be estimated by the sample standard deviation $s = \sqrt{\sum (x_i - \bar{x})^2/(n-1)}$. Then we would report the (estimated) standard error of $\bar{x}$ to be $s/\sqrt{n}$.

The idea of reporting a standard error is to give an indication of how much your estimate is expected to deviate from its target. The concept of confidence interval goes further and assigns a probability to the deviation – and allows you to vary the deviation until you get the amount of probability (commonly 0.90 or 0.95) that you want. If a 95% confidence interval is pretty short, then you can feel good that your estimate is on target. On the other hand, if an 80% confidence interval is quite wide, then your estimate may be imprecise.

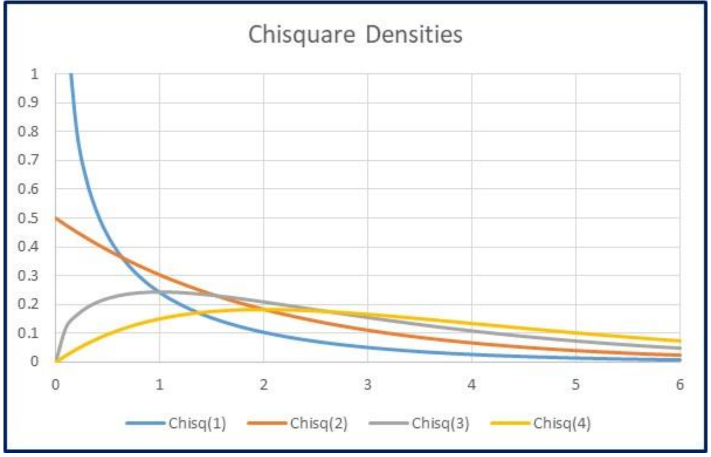
To develop the concept of confidence interval further, and to lay the foundation for hypothesis tests, we need to understand more about some of the distributions that are used to make confidence intervals and to test hypotheses. So next we turn to the chi-square, Student's *t*, and F distributions.

## Chi-square RVs.

Squaring a standardized normal RV gives a $\chi_1^2$ RV. We say "chi-square with one degree of freedom." Adding up the squares of *n* *independent* standardized normal RVs gives a $\chi_n^2$ RV. Here are the basic facts:

- Suppose *Z* is a RV with a standard normal distribution. Then $Z^2$ is a RV with a chi-square distribution with one degree of freedom.
- Suppose $Z_1, Z_2, ..., Z_n$ is a RS of standard normal RVs. Then $U = Z_1^2 + Z_2^2 + \cdots + Z_n^2$ is a RV that has a chi-square distribution with *n* degrees of freedom.
- A $\chi_n^2$ RV has a mean equal to *n* and a variance equal to 2*n*.

Figure 9. Some Chi-square densities



Ex: Suppose that $X_1, X_2, ..., X_n$ is a RS with common normal distribution having mean $\mu$ and variance $\sigma^2$. Then

- Each $\left(\frac{X_i - \mu}{\sigma}\right)^2$ is a $\chi_1^2$ RV.
- $\sum_{i=1}^n \left(\frac{X_i - \mu}{\sigma}\right)^2$ is a $\chi_n^2$ RV.
- $\left(\frac{\overline{X} - \mu}{\sigma / \sqrt{n}}\right)^2$ is a $\chi_1^2$ RV.
- $\frac{\sum_{i=1}^n (X_i - \overline{X})^2}{\sigma^2}$ is a $\chi_{n-1}^2$ RV. This is not immediately obvious, but follows from the mathematical identity $\sum_{i=1}^n \left(\frac{X_i - \mu}{\sigma}\right)^2 = \sum_{i=1}^n \left(\frac{(X_i - \overline{X}) + (\overline{X} - \mu)}{\sigma}\right)^2 = \frac{\sum_{i=1}^n (X_i - \overline{X})^2}{\sigma^2} + \left(\frac{\overline{X} - \mu}{\sigma / \sqrt{n}}\right)^2$. Now the left-hand side is $\chi_n^2$ and the second term on the right is $\chi_1^2$, so the first term on the right must be $\chi_{n-1}^2$. (This argument is not quite complete. The main piece missing is that the two terms on the right must be *independent* RVs – but that turns out to be true.)
- From the latter it follows that the sample variance, based on a normal ( $\mu, \sigma^2$ ) RS, is equal to $\frac{\sigma^2}{n-1}$ times a $\chi_{n-1}^2$ RV: $S^2 = \frac{\sum_{i=1}^n (X_i - \overline{X})^2}{n-1} = \frac{\sigma^2}{n-1} \frac{\sum_{i=1}^n (X_i - \overline{X})^2}{\sigma^2}$. Therefore, the expected value of $S^2$ is $E(S^2) = \frac{\sigma^2}{n-1} E\left(\frac{\sum_{i=1}^n (X_i - \overline{X})^2}{\sigma^2}\right) = \frac{\sigma^2}{n-1} E(\chi_{n-1}^2) = \frac{\sigma^2}{n-1}(n-1) = \sigma^2$. That is, the sample variance is an unbiased estimator of the population variance. Incidentally, this property accounts for why we divide by *n*-1 when calculating the sample variance. Dividing by *n*

would give the sample variance a slight downward bias as an estimator of the population variance.

Ex: It is true that the sample variance is an unbiased estimator of $\sigma^2$, even if the RS does not have a normal distribution.

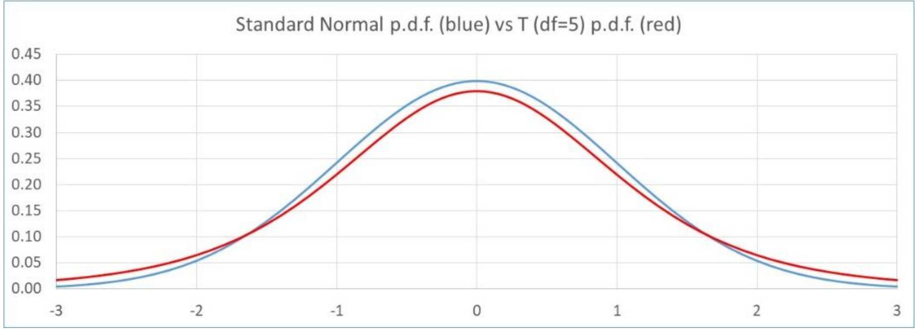
Ex: It is also true that the sample variance, based on a RS, is approximately equal to $\frac{\sigma^2}{n-1}$ times a $\chi^2_{n-1}$ RV if n is sufficiently large, even if the RS is not normal.

## Student *t* RVs.

Dividing a standard normal RV by the square root of the ratio of an ***independent*** $\chi^2$ to its degrees of freedom yields a Student *t* RV with the same number of degrees of freedom as the $\chi^2$ variable:

- Suppose *Z* is a RV with a standard normal distribution and *U* is an independent $\chi^2_n$ RV. Then $T = \frac{Z}{\sqrt{U/n}}$ is a Student *t* RV with *n* degrees of freedom.

Figure 10. T p.d.f. compared with standard normal p.d.f.



Ex: Suppose that $X_1, X_2$ is a RS with common normal distribution having mean 0 and variance 1. Then $T = \frac{X_1}{\sqrt{X_2{}^2/1}}$ has a Student *t* distribution with 1 degree of freedom.

Ex: Suppose that $X_1, X_2, ..., X_n$ is a RS with common normal distribution having mean $\mu$ and variance $\sigma^2$. Then $T = \frac{\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sigma^2}}{n-1}}} = \frac{\bar{X}-\mu}{S/\sqrt{n}}$ has a Student *t* distribution with *n*-1 degrees of freedom. This example illustrates the real explanation for why replacing an unknown population standard deviation $\sigma$ in $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$ with the known (but random) sample standard deviation $S$ to yield $\frac{\overline{X}-\mu}{S/\sqrt{n}}$ results in a Student *t* distribution: It is not actually a replacement of a known with an unknown – it is a more indirect division.

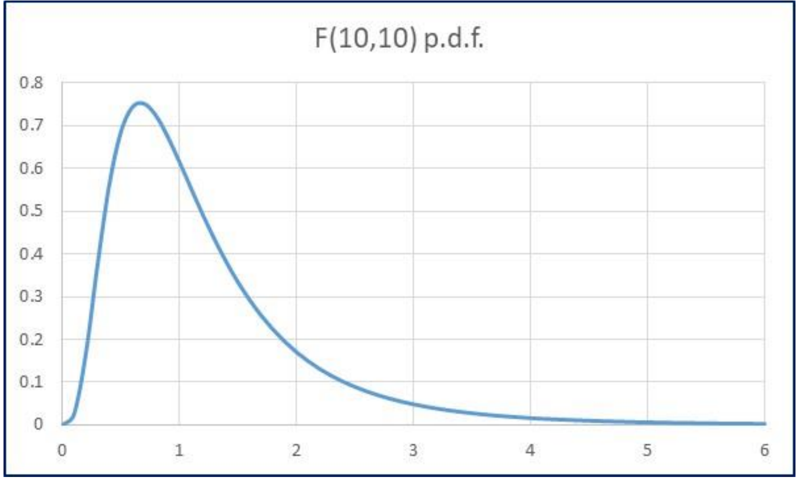
Ex: It is also true that $\frac{\overline{X}-\mu}{S/\sqrt{n}}$, based on a RS, has an approximate Student *t* distribution with *n*-1 degrees of freedom, if *n* is sufficiently large – even if the RS does not have a normal distribution.[14]

## F random variables.

Take the ratio of a $\chi^2$ RV to its degrees of freedom, and take the ratio of another $\chi^2$ RV to its degrees of freedom. If the two $\chi^2$ RVs are ***independent***, then the ratio of those two ratios has a F distribution:

- Suppose that $U$ is a RV with a $\chi_m^2$ distribution and $V$ is a RV with a $\chi_n^2$ distribution. If $U$ and $V$ are ***independent***, then $\frac{U/m}{V/n}$ is a RV that has an $F_{m.n}$ distribution.

Figure 11. F(10,10) p.d.f.



Ex: Suppose that $X_1, X_2$ is a RS with common normal distribution having mean 0 and variance 1. Then $T = \frac{X_1{}^2/1}{X_2{}^2/1}$ has an F distribution with 1, 1 degrees of freedom.

[14] In turn, $\frac{\overline{X}-\mu}{S/\sqrt{n}}$ is approximately standard normal if *n* is sufficiently large because $\frac{\overline{X}-\mu}{S/\sqrt{n}}$ converges in distribution to a standard normal RV by the Central Limit Theorem.

Ex: Suppose the normal RS model. Then $\left(\dfrac{\overline{X}-\mu}{\sigma/\sqrt{n}}\right)^2$ has a $\chi_1^2$ distribution, and $\dfrac{\sum_{i=1}^n (X_i-\overline{X})^2}{\sigma^2}$ has a $\chi_{n-1}^2$ distribution, and the two RVs are independent. Thus $\dfrac{\left(\dfrac{\overline{X}-\mu}{\sigma/\sqrt{n}}\right)^2/1}{\dfrac{\sum_{i=1}^n (X_i-\overline{X})^2}{\sigma^2}/(n-1)} = \dfrac{(\overline{X}-\mu)^2}{S^2/n}$ has an $F_{1,n-1}$ distribution. Note that this is the square of the RV $\dfrac{\overline{X}-\mu}{S/\sqrt{n}}$, which has the Student $t_{n-1}$ distribution.

With this background on the relationships among different distributions, I can now return to confidence intervals and then to hypothesis tests:

**Confidence intervals.**

If the endpoints of an interval are random variables, then we call the interval a **random interval**. The probability that the random interval contains a given constant *c* is called the **confidence coefficient** for *c*.

Ex: Suppose the normal RS model with common mean 0 and variance 1. Then $(\overline{X}-1.96/\sqrt{n},\ \overline{X}+1.96/\sqrt{n})$ is a random interval.

- The confidence coefficient for *c* = 0 is 0.95 because $P(\overline{X}-1.96/\sqrt{n}<0<\overline{X}+1.96/\sqrt{n}) = P(-1.96<\sqrt{n}\overline{X}<1.96)$ = 0.95 for all *n*. Note that $\sqrt{n}\overline{X} = \dfrac{\overline{X}-0}{1/\sqrt{n}}$ is N(0,1).
- The confidence coefficient for *c* = 1 is $P(\overline{X}-1.96/\sqrt{n}<1<\overline{X}+1.96/\sqrt{n})$ = $P(\sqrt{n}-1.96<\sqrt{n}\overline{X}<\sqrt{n}+1.96)$ = $\Phi\left(\sqrt{n}+1.96\right)-\Phi\left(\sqrt{n}-1.96\right)$, where $\Phi$ is the standard normal CDF (NORMSDIST in Excel). This does not have a constant value for all *n*, but is a decreasing function of *n*. E.g., it equals 0.8299 at *n* = 1, 0.1146 at *n* = 10, and 0.0060 at *n* = 20.

Ex: Suppose the normal RS model with common but unknown mean $\mu$ and variance 1. Then $(\overline{X}-1.96/\sqrt{n},\ \overline{X}+1.96/\sqrt{n})$ is a random interval. The confidence coefficient for *any* specific value of $\mu$ is 0.95 for all *n* because $P(\overline{X}-1.96/\sqrt{n}<\mu<\overline{X}+1.96/\sqrt{n}) = P(-1.96<\sqrt{n}(\overline{X}-\mu)<1.96)$ = 0.95.

A random interval that has the same confidence coefficient for all possible values of a parameter, as in the preceding example, is particularly useful in estimation. Only some random intervals have the same confidence coefficient for all possible values of a parameter. A **confidence interval** is a realization of a random interval that has the same confidence coefficient for all possible values of a parameter. So a confidence interval can be viewed as an estimate of an interval that contains a given probability.

A useful interpretation of a confidence interval, in general: Given a probability, like 0.95, and a random interval $(L,U)$ and a parameter $\theta$, $(L,U)$ is a 95% confidence interval for $\theta$ if 95% of the realizations $(l,u)$ of the random interval have the value of $\theta$ between the lower and upper limits: $l < \theta < u$, and 5% do not, *whatever the value of* $\theta$. A particular realization $(l,u)$ is also called a confidence interval, even though the particular realization either does or does not include $\theta$.

Ex: Suppose the normal RS model with common but unknown mean $\mu$ and known variance $\sigma^2$. Then ( $\bar{X} - 1.96\sigma/\sqrt{n}, \bar{X} + 1.96\sigma/\sqrt{n}$ ) is a random interval. The confidence coefficient for any specific value of $\mu$ is 0.95 because $P(\bar{X} - 1.96\sigma/\sqrt{n} < \mu < \bar{X} + 1.96\sigma/\sqrt{n}) = P(-1.96 < \dfrac{\bar{X} - \mu}{\sigma/\sqrt{n}} < 1.96)$ = 0.95. The realization $(\bar{x} - 1.96\sigma/\sqrt{n}, \bar{x} + 1.96\sigma/\sqrt{n})$ is a 95% confidence interval for $\mu$. It estimates the central interval that contains 95% of the probability of the distribution of $\bar{X}$. This example is rather unrealistic because one is unlikely to know the variance $\sigma^2$ if the mean $\mu$ is unknown.

In a possible abuse of terminology, sometimes the random interval is also called a confidence interval. E.g., in the preceding example, either ( $\bar{X} - 1.96\sigma/\sqrt{n}, \bar{X} + 1.96\sigma/\sqrt{n}$ ) or ( $\bar{x} - 1.96\sigma/\sqrt{n}, \bar{x} + 1.96\sigma/\sqrt{n}$ ) might be called a 95% confidence interval for $\mu$. However, only the former actually has 95% probability of containing $\mu$. The former is random; the latter is a realization.

Ex: Suppose the normal RS model with $n$ = 20, common but unknown mean $\mu$ and unknown variance $\sigma^2$. So this example differs from the preceding example mainly in having an unknown variance. Then ( $\bar{X} - 2.093S/\sqrt{n}, \bar{X} + 2.093S/\sqrt{n}$ ) is a random interval, where $S = \sqrt{\dfrac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}}$. The confidence coefficient for any specific value of $\mu$ is 0.95 because $P(\bar{X} - 2.093S/\sqrt{n} < \mu < \bar{X} + 2.093S/\sqrt{n}) = P(-2.093 < \dfrac{\bar{X} - \mu}{S/\sqrt{n}} < 2.093)$ = 0.95, since $\dfrac{\bar{X} - \mu}{S/\sqrt{n}}$ has a Student $t_{19}$ distribution. The realization ( $\bar{x} - 2.093s/\sqrt{n}, \bar{x} + 2.093s/\sqrt{n}$ ) is a 95% confidence interval for $\mu$, where $s = \sqrt{\dfrac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$.

Ex: A random sample of 60 apartments was selected from the population of all Austin apartments. Several variables were measured for each apartment. Among these variables was the monthly rent in dollars. The sample mean and sample standard deviation of the 60 rents were $\bar{y}$ = 572.27 and $s$ = 140.52. The model is the Random Sample model, with $X_1, X_2, ..., X_{60}$ independent with common distribution (not necessarily normal), having population mean $\mu$ and variance $\sigma^2$. Suppose the objective is to estimate the mean rent $\mu$ of the population. From previous discussion in these Notes, the best estimator from a variety of statistical perspectives is $\overline{Y}$, and its realization $\overline{y}$ = 572.27 is the best estimate of $\mu$. In estimating a parameter, one should not stop with the estimate, but should also report on the uncertainty of the estimate. That can be done by providing the standard error of the estimate or a confidence interval or both. The estimated standard error of the sample mean is $s/\sqrt{n}$ = 140.52 / $\sqrt{60}$ = 18.14. Based on the preceding examples, $(\overline{Y} - 2.001S/\sqrt{59}, \overline{Y} + 2.001S/\sqrt{59})$ is a random interval with approximate confidence coefficient 0.95 for $\mu$, using the T distribution with 59 degrees of freedom. The realization of this interval is $(572.27 - 2.001*140.52/\sqrt{59}, 572.27 + 2.001*140.52/\sqrt{59}$ ) = (535.97, 608.57), which therefore is an approximate 95% confidence interval for $\mu$. The random interval would be $(\overline{Y} - 1.96S/\sqrt{59}, \overline{Y} + 1.96S/\sqrt{59}$ ) if it used the normal instead of the T distribution. The realization with the normal form is (536.71, 607.82). There is little difference between the realizations of the T(59) interval and the normal interval. Thus, the simpler normal form may be used if the sample size is sufficiently large.

**Hypothesis tests.**

Recall that a statistical model is a set of specifications about the random variables that we observe or experiment with. In hypothesis testing, we try to sharpen the specification of the statistical model. We formulate two non-overlapping specifications, called the **null hypothesis** ($H_0$) and **alternative hypothesis** ($H_1$ - often denoted $H_a$) and ask the question, "Does the realization of the RVs provide sufficient evidence to reject the null hypothesis specification?" These specifications can be quite specific, such as $H_0: \mu = 0$ and $H_1: \mu = 1$. Or the specifications can be quite general, such as $H_0$: *the random variables are independent* and $H_1$: *the random variables are not independent*. Or the specifications can be partly specific, partly general, such as $H_0: \mu = 0$ and $H_1: \mu > 1$. Before performing the observation or experiment, we divide the population of outcomes for our RVs $X_1, X_2, ..., X_n$ into the **critical region** $C_R$ and the **acceptance region**. The critical region consists of all realizations $x_1, x_2, ..., x_n$ that we think provide sufficient evidence to reject $H_0$.[15] The acceptance region consists of all other possible realizations. We are free to choose a critical region and an acceptance region as we wish. But there are smart ways and foolish ways to do so. Hypothesis testing provides some smart ways. The central problem of hypothesis testing is how to choose a good critical region among the infinite number available.

So we examine the outcomes of the RVs in the observation or experiment and decide in favor of one or the other of the two specifications, depending upon where the realization lies. If in the critical region, then we decide in favor of $H_1$; if in the acceptance region, then we decide in favor of $H_0$ (or in the latter case, we withhold endorsement of $H_1$ - hypothesis testing is not quite even-handed in its treatment of the two hypotheses.) The **research hypothesis** represents

[15] As stated here, the critical region is vector-valued. However, most critical regions are defined in terms of one test statistic. That is OK – it amounts to the same thing. E.g., a critical region stated as $\overline{x} > 1.645$ means the same as all vector outcomes $(x_1, x_2, ..., x_n)$ whose mean > 1.645.

the specification whose truth we would like to establish. If we can, we generally set up the alternative hypothesis to be the research hypothesis. So we hope to find in favor of $H_1$. But the specification represented by the null hypothesis is accorded preferential treatment, in the sense that we give the benefit of the doubt to the null hypothesis in the event of an inconclusive test. We define our testing criteria so that only those realizations that are clear and convincing in favor of the alternative/research hypothesis result in rejection of the null hypothesis.

At first, this treatment may seem perverse. We want the alternative to be true, but we make it hard to reject the null. But it actually makes sense. The reason is that we want a strong endorsement of our research hypothesis. If we have bent over backward to give $H_0$ a leg up, and yet we still reject $H_0$, then we can feel pretty confident that $H_1$ is correct. On the other hand, because we agreed that we would count inconclusive outcomes as favoring the null hypothesis, acceptance of the null hypothesis is only a weak endorsement of the null hypothesis.

Here are the key points about hypothesis tests:

- A **statistical test of hypothesis** is a critical region $C_R$.
- Every possible critical region is a different test. So every different set of potential outcomes of the observable RVs is a different test.
- The actual decision to accept or to reject $H_0$ is made on the basis of the realization $(x_1, x_2, ..., x_n)$. If it falls into $C_R$, then reject $H_0$; otherwise, accept (fail to reject) $H_0$.[16]
- A **Type I error** is to reject $H_0$ when $H_0$ contains the true specification.
- A **Type II error** is to accept $H_0$ when $H_1$ contains the true specification.
- The **significance level** of the test is $\max\limits_{\text{all specificat ions in } H_0} P((X_1, X_2, ..., X_n) \in C_R)$ = $\max\limits_{\text{all specificat ions in } H_0} P(reject\ H_0)$. That is, you compute the probability of rejecting $H_0$ under each specification given by $H_0$, then take the maximum (supremum) of those probabilities. So the significance level is the maximum probability of committing a Type I error.
- The **power of the test at a given specification in** $H_1$ is $P((X_1, X_2, ..., X_n) \in C_R)$, computed at that specification. 1 - power is the probability of a Type II error at the specification under consideration.

## An Extended Example

A random sample of 60 apartments was selected from the population of all Austin apartments. Several variables were measured for each apartment. Among these variables was the monthly rent in dollars. The sample mean and sample standard deviation of the 60 rents were $\bar{y}$ = 572.27 and *s* = 140.52. The model is the Random Sample model, with $X_1, X_2, ..., X_{60}$ independent with common distribution (not necessarily normal), having population mean $\mu$ and variance $\sigma^2$.

[16] Because of the assignment of inconclusive outcomes to $H_0$, most careful researchers prefer to say “fail to reject $H_0$” instead of “accept $H_0$.”

- Suppose that having mean rent greater than 550 creates a financially enticing environment for investors. So we want to decide if the mean is greater than 550 or not. How should we choose which of the two possibilities to be the null and which the alternative hypothesis? One way to do it: $H_0 : \mu \le 550$ vs. $H_1 : \mu > 550$. With this set-up, a Type I error would be to conclude that the environment is enticing, when it really is not. The consequence of committing a Type I error would be to invest in an unpromising venture. A Type II error would be to conclude that the environment is unenticing, when really it is. The consequence of committing a Type II error would be to pass on investing in a promising venture. Since the null hypothesis ($H_0$) gets the benefit of the doubt, the environment is presumed to be unenticing unless the evidence is strong to the contrary. Such a stance toward the hypotheses would be considered *risk averse*.
- A test is determined by specification of a critical region $C_R$. Let us consider the test specified by $C_R = \{(x_1, x_2, ..., x_n); \bar{x} > 580\}$. That is, the hypothesis that the environment is unenticing is rejected if the sample mean rent exceeds 580. If the sample mean rent is large, then the population mean rent is also likely to be large, which makes for a favorable environment. But note that this test sets the cut-off point (**critical point**) higher than 550. This is because you need to allow some slack in order to be pretty sure that the true mean really does exceed 550. If the sample mean is between 550 and 580, it is a rather close call – and hypothesis testing assigns close calls to the null hypothesis.

What are the error probabilities of this test?

- Significance level = max P(Type 1 error) = max P(reject $H_0$ | $H_0$ is true) = $\max P(\bar{X} > 580 \,|\, \mu \le 550) = \max P\left(\frac{\bar{X} - \mu}{S/\sqrt{60}} > \frac{580 - \mu}{S/\sqrt{60}} \,|\, \mu \le 550\right)$. Now $\frac{\bar{X} - \mu}{S/\sqrt{60}}$ is $T_{59}$, which is approximately N(0,1), and the observed value of the standard error $S/\sqrt{60}$ is 18.14. So $\max P\left(\frac{\bar{X} - \mu}{S/\sqrt{60}} > \frac{580 - \mu}{S/\sqrt{60}} \,|\, \mu \le 550\right)$ is approximately $\max P\left(N(0,1) > \frac{580 - \mu}{18.14} \,|\, \mu \le 550\right)$. This probability is the area under the curve to the right of 580 when $\mu = 550$, as shown in Figure 12. For any value of $\mu$ less than 550 the picture would change by shifting the probability curve to the left. Clearly, such a shift would reduce the area to the right of 580. Therefore, the probability to the right of 580 for $\mu \le 550$ is clearly maximized when $\mu = 550$. Thus the significance level is approximately $P\left(N(0,1) > \frac{580 - 550}{18.14}\right) = P\left(N(0,1) > 1.65\right)$, which is approximately 0.05.

Figure 12. Significance level

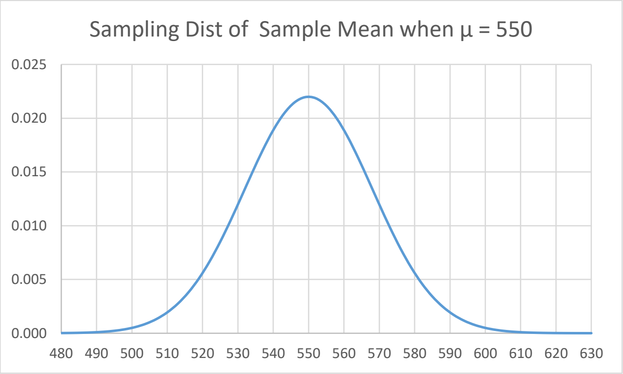
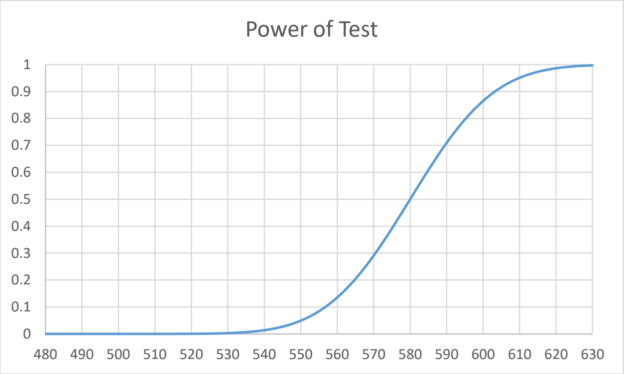


Figure 13. Power function



- Similarly, the power of this test is 1 – P(Type 2 error) = P(reject $H_0$ | $H_0$ is false) = $P(\overline{X} > 580 \,|\, \mu > 550) \;=\; P\left(\frac{\overline{X}-\mu}{S/\sqrt{60}} > \frac{580-\mu}{S/\sqrt{60}} \,|\, \mu > 550\right)$. Exactly as in the preceding bullet, this probability is approximately $P\left(N(0,1) > \frac{580-\mu}{18.14} \,|\, \mu > 550\right)$. This is a function of $\mu$ that has its minimum value for $\mu$ close to 550, where its value would be 0.05; the value climbs to 0.50 when $\mu$ = 580 and to 0.95 when $\mu$ = 610; the value continues to rise the larger $\mu$ grows. Figure 13 graphs the power as a function of $\mu$. The graph shows that it becomes easier to correctly reject $H_0$ the greater the difference between $\mu$ and $H_0$. The probability of a Type 2 error is 1 – this curve, as a function of $\mu$.

**The practical importance of significance level and power:** With your research hypothesis ensconced as the alternative hypothesis, you of course are rooting for rejection of $H_0$. But you want to be pretty sure that you are right if you reject $H_0$, so you want a low probability of falsely rejecting $H_0$. That is, you want a low significance level. If your sample size is set, you can make the significance level lower if you reduce the number of outcomes in $C_R$ – which makes it “harder” to reject $H_0$. For example, in the preceding Extended Example, you could increase the critical point and reject $H_0$ if $\overline{X} > 590$ instead of $\overline{X} > 580$. Unfortunately, reducing the size of $C_R$ makes it harder to reject $H_0$ whether or not $H_0$ is true. That is, if your research hypothesis is really correct, your reduction of $C_R$ in order to achieve a lower significance level also reduces your power, which is your ability to detect that your research hypothesis is correct. A trade-off between significance level and power is necessary. The standard approach to this trade-off is for you to choose a significance level that is the maximum that you can live with, then use the test (critical region) that has maximum power among all tests (critical regions) having your chosen significance level. So you want to use the most powerful test (critical region) possible. For example, as shown in the preceding Extended Example, the test $C_R$ = { $\overline{X} > 580$ } has significance level 0.05. It can be shown by advanced mathematical statistics that the power function of this test, shown in Figure 13, is greater than the power function of *any* other test that has significance level of 0.05. In most fields of science, the value 0.05 has become a kind of default criterion for the significance level of tests.

Having a **most powerful test** means that, among all tests of $H_0$ and $H_1$ that have a given significance level (like 0.05), no other test has greater power. Most of the hypothesis tests found in basic statistics textbooks are most powerful tests – or nearly so. Most powerful tests are very desirable, but they do not always exist. A most powerful test has minimum probability of a Type II error among all tests of a given significance level.

***p*-values.** When you compare two possible outcomes $(x_1, x_2, ..., x_n)$ and $(y_1, y_2, ..., y_n)$, you can usually say which one is more favorable for the alternative hypothesis.

Ex: If you test $H_0 : \mu \leq 0$ against $H_1 : \mu > 0$, an outcome with a sample mean of 2 is more favorable for $H_1$ than an outcome with a sample mean of 1.

Suppose that you test $H_0$ against $H_1$ and $(x_1, x_2, ..., x_n)$ is the realization of the RVs. Denote $C_{(x_1, x_2, ..., x_n)}$ the set of outcomes that are more favorable to $H_1$ than the observed realization $(x_1, x_2, ..., x_n)$. The **p-value** is $\max_{\text{all specifications in } H_0} P\{(X_1, X_2, ..., X_n) \in C_{(x_1, x_2, ..., x_n)}\}$. So the p-value is calculated the same as the significance level, but with a critical region that depends on the actual realization. This means that *the p-value is actually a statistic*. If $H_0$ is rejected, then the *p*-value cannot exceed the significance level. If $H_0$ is accepted, then the p-value must exceed the significance level. Thus, the decision to accept or reject $H_0$ can be based on the *p*-value instead of on the critical region: Reject $H_0$ if and only if *p*-value ≤ significance level. For reporting research results, *p*-values are preferred to significance levels. The reason is that reporting a *p*-value allows the reader to select his/her own significance level and make the decision: It is more informative to report that the *p*-value was 0.06 than to report that $H_0$ was rejected at a significance level of 0.10. Bayesians generally demur from calculating *p*-values and prefer other means to evaluate the relative truth of hypotheses.

Ex: In the preceding Extended Example, the observed realization $(x_1, x_2, ..., x_n)$ has $\bar{x}$ = 572.27 (and *s* = 140.52). Thus $C_{(x_1, x_2, ..., x_n)}$ = { $\bar{x}$ > 572.27}. The p-value is then

$\max P(\bar{X} > 572.27 \mid \mu \leq 550)$ = $\max P\left(\frac{\bar{X} - \mu}{S / \sqrt{60}} > \frac{572.27 - \mu}{S / \sqrt{60}} \mid \mu \leq 550\right)$ approx. =

$\max P\left(N(0,1) > \frac{572.27 - \mu}{18.14} \mid \mu \leq 550\right)$ = $P\left(N(0,1) > \frac{572.27 - 550}{18.14} = 1.2274\right)$ = 0.1098. If we use the test $C_R$ = { $\bar{X} > 580$ }, which has significance level 0.05, then we would not reject $H_0$ both because the observed sample mean 572.27 < 580 (the critical point) and because the p-value 0.1098 > 0.05 (the significance level).

Ex: Suppose the normal RS model with common but unknown mean $\mu$ and known variance $\sigma^2$, and we test $H_0 : \mu \leq c$ against $H_1 : \mu > c$. Define the critical region to be $C_R$ =

$\{(x_1,x_2,...,x_n);\ \bar{x} > c + 1.645\sigma/\sqrt{n}\}$ (equivalently, CR = $\{\bar{x};\ \frac{\bar{x}-c}{\sigma/\sqrt{n}} > 1.645\}$ ). Then the significance level is $\max_{\text{all specifications in } H_0} P((X_1,X_2,...,X_n) \in C_R)$ = $\max_{\mu \le c} P(\ \bar{X} > c + 1.645\sigma/\sqrt{n})$ = $\max_{\mu \le c} P\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} > \frac{c-\mu}{\sigma/\sqrt{n}} + 1.645\right)$. Since $c - \mu \ge 0$ in $H_0$, then the probability is maximized if we set $c = \mu$ (Why?). So the significance level is $P\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} > 1.645\right)$ = .05. For a given $\mu > c$, the power is $P((X_1,X_2,...,X_n) \in C_R)$ = $P(\ \bar{X} > c + 1.645\sigma/\sqrt{n})$ = $P\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} > \frac{c-\mu}{\sigma/\sqrt{n}} + 1.645\right)$ = $P(\ Z > \frac{c-\mu}{\sigma/\sqrt{n}} + 1.645)$, where *Z* is standard normal. If $\mu$ is only a little above *c*, then the power is very small – only a little above 0.05. This means that the ability of the test to decide in favor of the research hypothesis is weak when truth is only a little different from the null hypothesis. But as the separation increases and $c - \mu$ widens and becomes more negative ($c - \mu < 0$ in $H_1$), then the power increases. When the gap between *c* and $\mu$ reaches -3.29 standard errors, then the power equals 0.95. (Why?) As the sample size increases, the power also increases. This test is *uniformly* most powerful, i.e., it is the most powerful test regardless of the values of $\mu$ being tested. Suppose *c* = 0, *n* = 16, *σ* = 2, and the realization $(x_1,x_2,...,x_n)$ has $\bar{x}$ = 0.75. Then $\frac{\bar{x}-\mu}{\sigma/\sqrt{n}} = \frac{.75-0}{2/\sqrt{16}} = 1.5$, which is not in the critical region. Therefore, we accept (fail to reject) $H_0$. Realizations with standardized sample means that exceed 1.5 are more favorable to $H_1$ than realizations with standardized sample mean of 1.5 are. So the p-value is $P\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} > 1.5\right)$ = 0.0668.

Ex: Suppose the normal RS model with common but unknown mean *μ* and *un*known variance $\sigma^2$, and we test $H_0: \mu \le c$ against $H_1: \mu > c$. (So this example is the same as the previous example, except that the variance is now unknown.) Further suppose that *n* = 16. Define the critical region to be CR = $\{(x_1,x_2,...,x_n);\ \bar{x} > c + 1.753s/\sqrt{n}\}$ (equivalently, CR = $\{\bar{x};\ \frac{\bar{x}-c}{s/\sqrt{n}} > 1.753\}$). Then the significance level is $\max_{\mu \le c} P(\ \bar{X} > c + 1.753S/\sqrt{n})$ = $\max_{\mu \le c} P\left(\frac{\bar{X}-\mu}{S/\sqrt{n}} > \frac{c-\mu}{S/\sqrt{n}} + 1.753\right)$. Since $c - \mu \ge 0$ in $H_0$, then the probability is maximized if we set $c = \mu$. (Why?) So the significance level is $P\left(\frac{\bar{X}-\mu}{S/\sqrt{n}} > 1.753\right)$ = .05, from Student's *t* distribution. For a given $\mu > c$, the power is $P(\ \bar{X} > c + 1.753S/\sqrt{n})$ =

$P\left(\frac{\overline{X}-\mu}{S/\sqrt{n}} > \frac{c-\mu}{S/\sqrt{n}} + 1.753\right) = P(T_{n-1} > \frac{c-\mu}{S/\sqrt{n}} + 1.753)$. This probability is not straightforward to calculate, since there is a RV on the right-hand side of the inequality (namely, *S*) and since $T_{n-1}$ and *S* are dependent. But the probability can be approximated by pretending that the sample realization *s* of *S* is $\sigma$. The justification for this pretense is that *S* converges to $\sigma$.[17] If the sample size is small, there may be considerable error in the approximation, but in large samples, the error is likely to be small. As in the previous example, if $\mu$ is only a little above *c*, then the power is very small – only a little above 0.05. The ability of the test to decide in favor of the research hypothesis is weak when truth is only a little different from the null hypothesis. But as the separation increases and $c-\mu$ widens and becomes more negative ($c-\mu<0$ in $H_1$), then the power increases. When the gap between *c* and $\mu$ reaches -3.506 (estimated) standard errors, then the power is approximately 0.95. (Why?) As the sample size increases, the power also increases. Suppose *c* = 0, *n* = 16, *s* = 3, and the realization $(x_1, x_2, ..., x_n)$ has $\bar{x}$ = 0.75. Then $\frac{\bar{x}-\mu}{s/\sqrt{n}} = \frac{.75-0}{3/\sqrt{16}} = 1.0$, which is not in the critical region. Therefore, we accept (fail to reject) $H_0$. Realizations with standardized sample means that exceed 1.0 are more favorable to $H_1$ than realizations with standardized sample mean of 1.0 are. So the p-value is $P\left(\frac{\overline{X}-\mu}{S/\sqrt{n}} > 1.0\right)$ = 0.1666 (from Student *t* distribution with 15 degrees of freedom.)

Ex: Suppose the RS model with common but unknown mean $\mu$ and *un*known variance $\sigma^2$, and we test $H_0: \mu \le c$ against $H_1: \mu > c$. (So this example is the same as the previous example, except that, in addition to the variance being unknown, the RVs are not necessarily normal.) In this case, everything in the previous example is approximately true, provided the sample size is sufficiently large.

[17] *S* converges to $\sigma$ almost surely, in MSE, and in probability.