Statistical Inference (Conover Chapter Two)

STAT 345-01: Nonparametric Statistics [*]

Fall Semester 2018

Contents

Tuesday 28 August 2018 – Read Chapter 1 of Conover; refer to Chapter 1 of Hollander

0 Preliminaries

0.1 Administrata

- Introductions!
- Syllabus
- Instructor's name (Whelan) rhymes with "wailin'".
- Text: Conover, *Practical Nonparametric Statistics*, 3rd edition.
- Recommended text: Hollander, Wolfe and Chicken, *Nonparametric Statistical Methods*, 3rd edition.
- Other useful books:
  - Higgins, *Introduction to Modern Nonparametric Statistics*, 1st edition.
  - Gibbons and Chakraborti, *Nonparametric Statistical Inference*, 5th edition.
- Course website: `http://ccrg.rit.edu/~whelan/STAT-345/`
  - Will contain links to notes and problem sets; course

[*] Copyright 2018, John T. Whelan, and all that

calendar is probably the most useful.
  - Course calendar: *tentative* timetable for course.

- Course work:
  - Please read the relevant sections of the textbook *before* class so as to be prepared for class discussions.
  - Conover has many short exercises, and the answers to the odd numbered ones are in the book. You should do as many of these as you can as you go along, to check that you understand how things work.
  - There will be quasi-weekly homeworks. Collaboration is allowed an encouraged, but please turn in your own work, as obviously identical homeworks may not receive credit.
  - We'll have a longer-term project towards the end of the semester.
  - There will be two prelim exams, in class, and one cumulative final exam.

- Grading:

25%	Problem Sets, Including Project
20%	First Prelim Exam
20%	Second Prelim Exam
35%	Final Exam

You'll get a separate grade on the "quality point" scale (e.g., 3.1667–3.5 is the B+ range) for each of these five components; course grade is weighted average.

0.2 Outline

1. Review/Basics of Probability and Statistical Inference (Chapters One and Two)
2. Binomial Tests (Chapter Three)
3. Rank-Based Tests (Chapter Five)
4. Kolmagorov-Smirnov Statistics (Chapter Six)
5. Contingency Tables (Chapter Four)

0.3 Perspective on Nonparametric Methods

In your introductory statistics course (MATH-252 or STAT-205), you learned about an array of statistical procedures for estimating quantities and testing hypotheses. Many of them were based on the properties of the normal distribution. They probably seemed arbitrary at the time, but as you'll learn if you go on in statistics, you can often define so-called **optimal procedures** which are known to outperform all the alternatives if the underlying properties of the random data are known, e.g., if you have a sample drawn from a specified distribution with unknown parameters. On the other hand, in this course, we'll be concerned with methods which work reasonably well even when you know little to nothing about the underlying distribution. To paraphrase former US Defense Secretary Donald Rumsfeld, parametric statistical methods deal with "known unknowns" (unknown parameters in a known distribution), while nonparametric methods are designed for "unknown unknowns" (situations where you don't even know the family of distributions you're dealing with).

Note that "nonparametric" is often something of a misnomer; if we're trying to estimate some quantity, like the median of a distribution, you could consider that to a parameter. A more general term for the kind of methods we'll look at is **robust**, which means that while they may not be the most efficient under ideal circumstances, they still perform well when simplifying assumptions are violated.

1 Basics of Probability

You should look over your notes from MATH-251 or STAT-205 on the details of probability theory; there is a brief review in Chapter 1 of Conover.[1] Very briefly, a probability $P(A)$, where $0 \leq P(A) \leq 1$ can be assigned to an event A. An event is, in general, a statement which can be either true or false. In the classical or **frequentist** formulation of probability, it must be the random outcome of a repeatable experiment. If we repeat the experiment N times, and N_A is the number of times that A turns out to be true, then N_A/N should approach $P(A)$ as N becomes large, i.e., $\lim_{N\to\infty} \frac{N_A}{N} = P(A)$. In the more general **Bayesian** formulation of probability, A can refer to any true-false proposition, and then $P(A|I)$ represents a degree of certainty, given the available information I, that A is true.

Events can be combined in various ways, but for our purposes we'll be interested in AB (also referred to as $A \cap B$, A and B, $A \wedge B$, or A, B), which is true if *both* A and B are true, and false otherwise, and define $P(AB)$ as the probability of this. We can also define the conditional probability that A is true given the assumption that B is true,

$$P(A|B) = \frac{P(AB)}{P(B)} \tag{1.1}$$

We say that the events A and B are **independent** if $P(AB) = P(A)P(B)$, which implies that $P(A|B) = P(A)$ and $P(B|A) = P(B)$.

1.1 Probability Distributions

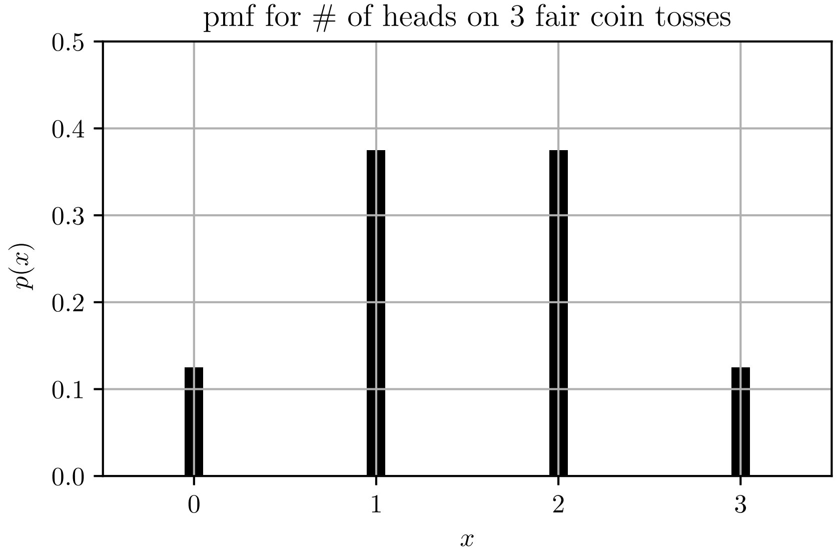
A **random variable** is a quantity X whose value is not known, but is described using probabilities. The most general way to talk about a random variable is via its cumulative distribution function (cdf) (Conover calls this the "distribution function")

$$F(x) = P(X \leq x) \tag{1.2}$$

A discrete random variable can only take on certain specific values, so we can describe it using a probability mass function (pmf) (Conover calls this a "probability function" and, confusingly, uses the notation $f(x)$)

$$p(x) = P(X = x) \tag{1.3}$$

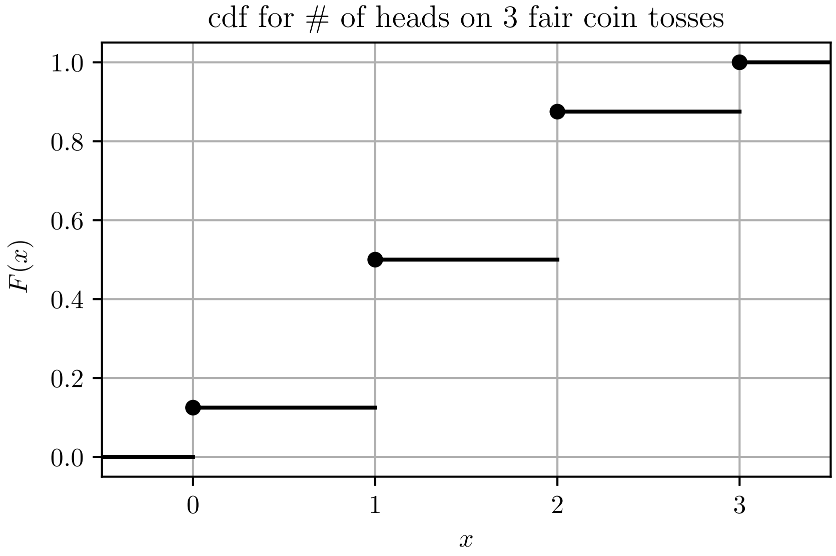
An example is the number of heads in three tosses of a fair coin, which has a pmf[2] that looks like this:



and a cdf that looks like this:

[1] My most recent analogous notes are at http://ccrg.rit.edu/~whelan/1016-345/

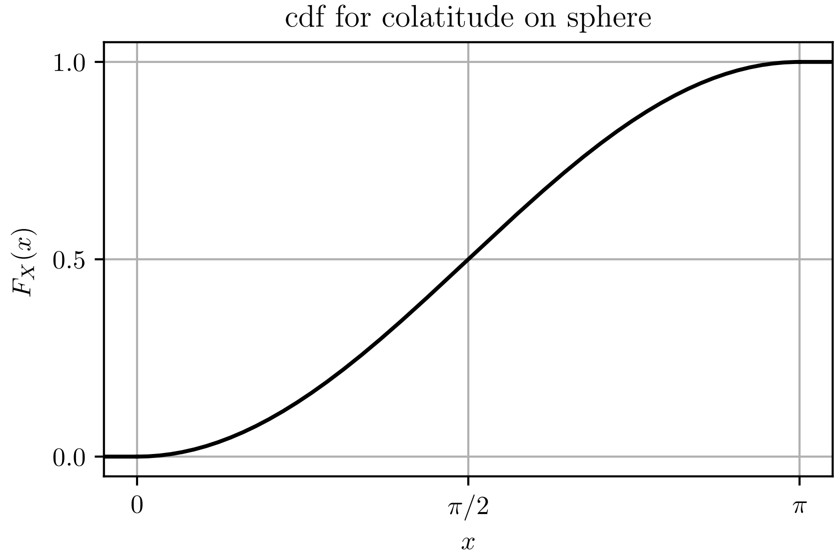
[2] This is a binomial random variable with $n = 3$ trials and a probability of $p = 0.5$. The general form of the pmf is $p(x) = \frac{n!}{(n-x)!x!} p^x (1-p)^{n-x}$, $x = 0, 1, \ldots, n$. We'll be using this distribution a lot starting next week.



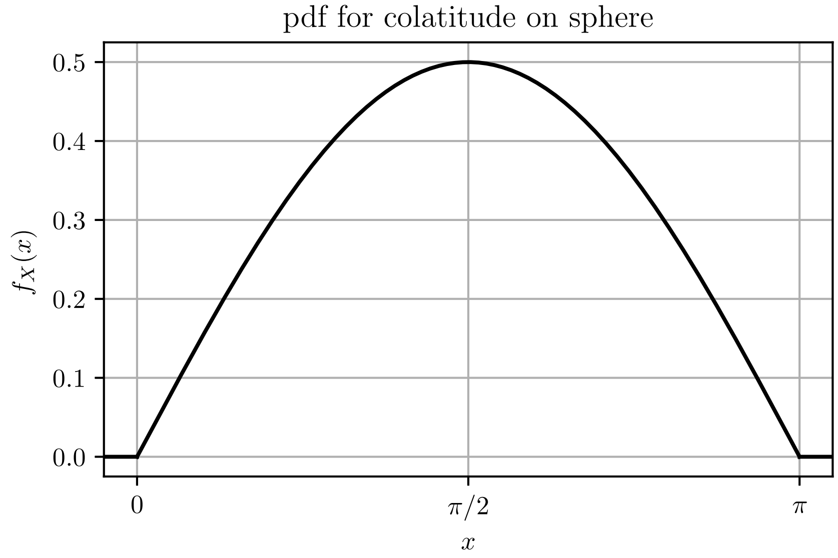
At each of the possible values of the random variable, the cdf $F(x)$ jumps by an amount equal to the pmf $p(x)$. If we want the probability for a discrete random variable to lie between two values a and b, we can add up all of the probabilities for values in between the two

$$P(a \leq X \leq b) = \sum_{x=a}^{b} p(x) = F(b) - \lim_{\epsilon \to 0} F(a - \epsilon) \qquad (1.4)$$

On the other hand, a continuous random variable has a continuous cdf $F(x)$ like this



Its derivative $f(x) = F'(x)$ is called the probability density function (pdf), like this



The pdf can be used to define the probability that X lies in a certain interval:

$$P(a<X<b) = P(a<X\leq b) = F(b)-F(a) = \int_a^b f(x)\,dx \quad (1.5)$$

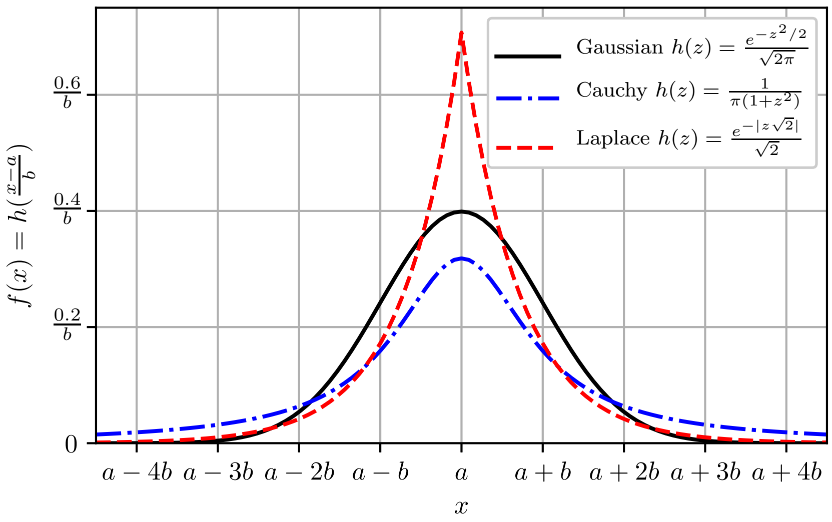
A commonly used continuous distribution is the normal, aka Gaussian, distribution

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\,e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (1.6)$$

This distribution has parameters μ and σ. It's an example of a distribution family with a location parameter $a=\mu$ and a scale parameter $\beta=\sigma$. In general, we can write such pdf as

$$f(x) = \frac{1}{b}\,h\left(\frac{x-a}{b}\right) \quad (1.7)$$

where $h(z)$ is some function that doesn't involve a or b. Some examples of these families are plotted below:



Of course, distributions may have additional parameters which influence them in more complicated ways, generally known as **shape parameters**, for example the number of degrees of freedom ν in a Student-t distribution, which has a pdf of the form

$$h(z) \propto \frac{1}{\left(1+\frac{z^2}{\nu}\right)^{\frac{\nu+1}{2}}} \quad (1.8)$$

(We've omitted the form of the normalization constant, which is not particularly enlightening.) When $\nu=1$ this is the Cauchy distribution shown above, and when ν becomes large, it is close to the normal distribution.

1.2 Quantiles

An important quantity associated with a probability distribution is the **quantile** associated with a value p such that $0\leq p\leq 1$. The pth quantile of a distribution is the value x_p such that the corresponding random variable has a probability p of lying below x_p. Stated more precisely (which is important for discrete distributions),

$$P(X<x_p)\leq p \qquad \text{and} \qquad P(X>x_p)\leq 1-p \quad (1.9)$$

The 0.5 quantile is also known as the median or the 50th percentile. The 0.05 quantile is the 5th percentile, the 0.75 quantile is the 75th percentile or the third quartile, etc.

Note that for a continuous random variable with pdf $f(x)$, the p quantile can be defined indirectly as

$$\int_{-\infty}^{x_p} f(x)\,dx = F(x_p) = p \quad (1.10)$$

1.2.1 Median vs Mean

The median is a measure of the location of a distribution. Another measure is the mean, or expectation value, which is the average of the distribution, whose form depends on whether the distribution is discrete or continuous:

$$\mu = E\left[X\right] = \int_{-\infty}^{\infty} x\, f(x)\, dx \qquad \text{or} \qquad \sum_{x} x\, p(x) \tag{1.11}$$

However, the median is sometimes a more useful generic property than the mean. For example, for the Cauchy distribution, the mean is not defined because the integral is the sum of two infinite contributions, one positive and one negative:

$$\begin{aligned}\int_{-\infty}^{\infty} \frac{x\,dx}{\pi(1+x^2)} &= \int_{-\infty}^{0} \frac{x\,dx}{\pi(1+x^2)} \int_{0}^{\infty} \frac{x\,dx}{\pi(1+x^2)} \\ &= -\lim_{A\to\infty} \frac{\ln(1+A^2)}{2\pi} + \lim_{B\to\infty} \frac{\ln(1+B^2)}{2\pi}\end{aligned} \tag{1.12}$$

On the other hand, the median is zero (or more generally the location parameter a), since

$$\int_{-\infty}^{0} \frac{dx}{\pi(1+x^2)} = \int_{0}^{\infty} \frac{dx}{\pi(1+x^2)} = \frac{1}{2} \tag{1.13}$$

Thursday 30 August 2018 – Read Sections 2.1-2.3 of Conover; refer to Chapter 1 of Hollander

1.3 Using Software to “Look Up” Properties of Standard Distributions

The appendices of statistics books are filled with tables of values and probabilities for various distributions. This is rapidly becoming an anachronism, like trigonometry books before the advent of calculators which were full of tables of sines and cosines. Now we have statistical software which can return all the desired probabilities and percentiles for most standard distributions. The statistical computing language R of course has many of these functions built in, but in this class I’ll show you some of the corresponding tools available in Python thanks to the `scipy.stats` package from Scientific Python. We can fire up an interactive session with[3] `ipython --pylab` and try a few commands:

```
from __future__ import division
import numpy as np
from scipy import stats
stats.norm.cdf(1.)
stats.norm.sf(1.)
stats.norm.cdf(2.)
stats.norm.cdf(1.,scale=0.5)
stats.norm.cdf(0.5,scale=0.5)
stats.norm.cdf(1.3,scale=0.5,loc=0.3)
x = np.linspace(-5,5,100)
print(x)
Fx = stats.norm.cdf(x)
from matplotlib import pyplot as plt
plt.figure()
plt.plot(x,Fx)
plt.title(’Standard normal cdf’)
plt.savefig(’normalcdf.eps’,bbox_inches=’tight’)
plt.figure()
fx = stats.norm.pdf(x)
plt.plot(x,fx)
plt.title(’Standard normal pdf’)
```

[3] The `--pylab` imports a bunch of numpy and matplotlib functions into the namespace; it has the same effect as `from pylab import *`, which is kind of a crutch, and I prefer to actually access numpy functions explicitly. It does make plotting work more smoothly.

Documentation is available via `https://docs.scipy.org/doc/scipy/reference/stats.html` . In a nutshell, I can get something like the cdf at $x = 1.5$ for the normal distribution with $\mu = 1$ and $\sigma = 2$ with `stats.norm.cdf(1.5,loc=1.,scale=2.)` or equivalently `stats.norm(loc=1.,scale=2.).cdf(1.5)`. (I can either specify the parameters of the distribution when I'm invoking it, or in the argument of the request for the cdf.) In place of `norm` I can ask for other distributions like `t`, `gamma`, `chi2`, etc. (See the documentation for the full list.) Instead of `cdf(x)` for the cdf $F(x)$, we could also use `pdf(x)` for the pdf $f(x)$, `sf(x)` for the "survival function" $S(x) = 1-F(x)$ (this is useful if we're out on the tail so $F(x)$ is close to 1), `ppf(p)` for the quantile x_p, `isf(p)` for the inverse survival function x_{1-p}, and various other possibilities. If you look up the documentation for a particular distribution, you'll get a list of the possible methods. As a bit of jargon, the distribution in this construction is called an object and the function you're ultimately calling is a method on that object.

2 Statistical Inference

2.1 Random Samples

A typical scenario[4] in statistical inference involves a data set $x_1, x_2, \ldots, x_n \equiv \{x_i\} \equiv \mathbf{x}$ which is assumed to be a realization of n independent random variables $X_1, X_2, \ldots, X_n \equiv \{X_i\} \equiv \mathbf{X}$ all drawn from some distribution $f(x)$. In the formalism of probability theory, there is a joint distribution function

$$f(x_1, x_2, \ldots, x_n) = f(x_1)f(x_2)\cdots f(x_n) \ . \tag{2.1}$$

the typical goal of statistical inference is to say something about the distribution function $f(x)$ based on the observed data $\{x_i\}$.[5] In conventional applications, this usually means constraining the unknown values of parameters $\theta_1, \theta_2, \ldots, \theta_p \equiv \{\theta_j\} \equiv \boldsymbol{\theta}$ in the distribution $f(x;\boldsymbol{\theta})$, such as the mean μ and/or standard deviation σ in a normal distribution. Of course, in non-parametric statistics, the information we're interested in not typically parameter values but rather more general information about the sampling distribution $f(x)$, but it's helpful to have a reminder about the standard procedures.

If our uncertainty about the sampling distribution $f(x;\theta)$ can be described by a parameter θ, the standard problem of parametric inference is to make a statement about the unknown value θ given the actual observed data $\mathbf{x} \equiv x_1, x_2, \ldots, x_n$. The difficulty is that what we actually have a mathematical description for is the probability distribution of the random vector $\mathbf{X}$ given the value of θ: $f(\mathbf{x};\theta)$. This tells us about the probabilities associated with collecting additional data sets of the same sort as $\mathbf{x}$. Bayesian statistics gets around this by interpreting $f(\mathbf{x};\theta)$ as a conditional probability distribution $f(\mathbf{x}|\theta,I)$ (given a value of θ and background information I, e.g., that the parameterized model is the correct description in the first place) and using Bayes's Theorem to construct

$$f(\theta|\mathbf{x},I) = \frac{f(\mathbf{x}|\theta,I)\,f(\theta|I)}{f(\mathbf{x}|I)} \tag{2.2}$$

which is a **posterior probability distribution** describing our knowledge of the parameter θ after we've collected the data $\mathbf{x}$.

Classical frequentist methods instead construct a statistic $T(\mathbf{x})$ from the data, and describe the probability distribution

[4] There are of course more complicated scenarios, like two random samples $\{X_i\}$ and $\{Y_j\}$ from different distributions, but we'll consider the simple case first for convenience.

[5] When talking about the distribution $f(x)$, it's conventional to talk about a random variable X with that distribution. Of course X_1, X_2, etc all have this distribution, so statements about e.g., $E\left[X\right]$ apply equally well to $E\left[X_1\right]$, $E\left[X_2\right]$, etc.

of the random variable $T(\mathbf{X})$ for possible values of θ. Roughly speaking, reasonable values of θ are those for which the value $T(\mathbf{x})$ is a "typical" value according to the behavior of the statistic $T(\mathbf{X})$.

2.2 Confidence Intervals

To give a concrete example, suppose $f(x)$ is a distribution with mean $\mu = E[X]$ and variance $\operatorname{Var}(X) \equiv E[(X-\mu)^2] = \sigma^2$. A standard result from intro statistics says that if we construct the sample mean

$$\overline{X} = \frac{X_1 + \cdots + X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{2.3}$$

and sample variance

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 \tag{2.4}$$

then the statistics have properties

$$E\left[\overline{X}\right] = \mu \qquad \text{and} \qquad \operatorname{Var}(\overline{X}) = \frac{\sigma^2}{n} \tag{2.5}$$

and

$$E\left[S^2\right] = \sigma^2 \tag{2.6}$$

Note that $\overline{X}$ is known as an **unbiased estimator** of μ because $E\left[\overline{X}\right] = \mu$, and likewise for S^2 with σ^2. We can use this to construct statistics

$$Z = \frac{\overline{X} - \mu}{\sqrt{\sigma^2/n}} \tag{2.7}$$

and

$$T = \frac{\overline{X} - \mu}{\sqrt{S^2/n}} \tag{2.8}$$

Some important results from introductory statistics are that

1. When the underlying distribution $f(x;\mu,\sigma)$ is a Gaussian (normal) $N(\mu.\sigma^2)$ with the appropriate parameters, then Z is a standard normal random variable $N(0,1)$, and T is Student-t distributed with $n-1$ degrees of freedom. (This is part of Student's Theorem.)
2. When the sample size is large enough, for any underlying distribution, both Z and T are approximately normally distributed. (This is an application of the Central Limit Theorem.) "Large enough" typically means $n \gtrsim 30$ for Z and $n \gtrsim 40$ for T.

This can be used to construct a *confidence interval* for the parameter μ. For example, if n is large, the statistic T will have a 5% chance of exceeding the 95th percentile of the standard normal distribution[6] $z_{.95} \approx 1.645$. This means that

$$.05 \approx P\left(\frac{\overline{X} - \mu}{\sqrt{S^2/n}} > z_{.95}\right) = P\left(\overline{X} - z_{.95}\sqrt{S^2/n} < \mu\right) \tag{2.9}$$

Likewise, it has a 5% chance of being less than the 5th percentile $z_{.05} \approx -1.645$. (Note that $z_{.05} = -z_{.95}$ because the standard normal pdf is symmetric about the origin.) So we have

$$.05 \approx P\left(\frac{\overline{X} - \mu}{\sqrt{S^2/n}} < z_{.05}\right) = P\left(\mu < \overline{X} - z_{.05}\sqrt{S^2/n}\right) \tag{2.10}$$

This means that there is a 90% chance that the interval bounded by $\overline{X} \pm z_{.95}\sqrt{S^2/n}$ will contain the true mean value μ. We call this a 90% confidence interval. Note that the random quantity associated with the probability is not the unknown value of μ (which is treated as fixed but unknown in the classical frequentist formalism), but rather the endpoints of the confidence

[6] Note that this is the opposite of the notational convention in Devore, where z_α is the $(1-\alpha)\times 100$th percentile rather than the $\alpha \times 100$th.

interval constructed from random data. Given an actual data set $\{x_i\}$, we construct the confidence interval as $\overline{x} \pm z_{.95} s/\sqrt{n}$.

One thing to note is that the central limit theorem means that the confidence interval construction is correct even if the sampling distribution is not Gaussian, as long as it has a finite mean μ and variance σ. However, it may not be the narrowest confidence interval we could construct at that confidence level, if the underlying distribution is non-Gaussian.

2.3 Empirical Distributions

One piece of standard descriptive statistics that can actually be considered as a form of nonparametric inference is the histogram, which can be thought of as an approximation to either the pdf $f(x)$ or the pmf $p(x)$, depending on the sort of random variable we're dealing with.

```
from __future__ import division
import numpy as np
from scipy import stats
mydist = stats.gamma(3,scale=10)
x_i = mydist.rvs(size=20)
hist(x_i,color='w',edgecolor='k',normed=True)
x = np.linspace(0.,90.,1000)
plot(x,mydist.pdf(x))
```

One obvious issue with this is the arbitrariness in binning the histogram. A way around that is to estimate the cdf $F(x) = P(X \le x)$ rather than the pdf or pmf. The empirical distribution function $\hat{F}(x;\{x_i\})$ (which Conover calls $S(x)$) is just the fraction of observations $\{x_i\}$ that are less than or equal to x. This can be easily estimated by using NumPy's Boolean array construction; `A<=B` is an array containing `True` wherever the inequalities is satisfied, and `False` wherever it's not. If we take the mean of this array, it gives the fraction of true values.

```
mymask = x_i[None,:] <= x[:,None]
mymask
mymask.shape
Phat = np.mean(mymask,axis=-1)
figure()
plot(x,Phat)
plot(x,mydist.cdf(x))
```

Conover considers a more involved method of estimating the survival function $1 - F(x)$ in cases where some of the data are missing, known as the Kaplan-Meier estimator, but we'll skip over that as it's a bit advanced for our purposes right now.

Tuesday 4 September 2018 – Read Sections 2.4-2.5 of Conover; refer to Chapter 1 of Hollander

2.4 Hypothesis Tests

Last week we considered confidence intervals, which are a form of parameter estimation. Another major sort of statistical inference, and one which can be easily extended to nonparametric scenarios, is **hypothesis testing**. Given a set of data $\mathbf{x} \equiv x_1, \ldots, x_n \equiv \{x_i\}$, we wish to distinguish between two competing statements about the probability distribution $f(\mathbf{x})$ describing the random vector $\mathbf{X}$ of which $\mathbf{x}$ is supposed to be an instance. We call them the **null hypothesis** H_0 and the **alternative hypothesis** H_1. In a Bayesian approach we would make some comparison between posterior probabilities $P(H_0|\mathbf{x})$ and $P(H_1|\mathbf{x})$, but in the classical formulation we have to make some indirect statement involving the possible probability distributions for $\mathbf{X}$.

Classical hypothesis testing treats the two hypotheses differently. The null hypothesis H_0 tends to describe the absence of some effect which is present in H_1. A hypothesis test is a rule for choosing between two alternatives given the observed data $\mathbf{x}$, but it's not as simple as "pick H_0" or "pick H_1". Rather, the two possibilities are

1. Reject H_0 (in favor of H_1).
2. Don't reject H_0.

At no point do we actually *accept* either hypothesis. And in particular, a negative test result doesn't mean we rule out the effect described by H_1; it might be that the data just don't contain enough information to see it. As a matter of terminology, the set of all points in the n-dimension "sample space" (whose coordinates are $(x_1, \ldots, x_n)$) for which the test says to reject H_0 is known as the **critical region** or **rejection region**.

To give a concrete example, consider a sample of size n drawn from some distribution with a finite mean μ and variance σ. Let H_0 specify that $\mu = 0$ and H_1 that $\mu > 0$. We define a test that rejects H_0 if

$$z = \frac{\overline{x}}{\sigma/\sqrt{n}} > 1.645 \tag{2.11}$$

(where $\overline{x}$ is the sample mean of the data and we assume both hypotheses specify the value σ of the population standard deviation) and fails to reject if $z \leq 1.645$. The value z is a realization of the random variable

$$Z = \frac{\overline{X}}{\sigma/\sqrt{n}} \tag{2.12}$$

which is known as a **test statistic**.

2.4.1 Significance

If there were no uncertainty or randomness, the outcome of a hypothesis test would be definitive. If H_1 were true, the test would reject H_0, and if H_0 were true, the test would not reject it. But of course, there is a chance that the test will give the "wrong" answer. Rejecting H_0 if it's true is known as a **Type I Error** or a false alarm. Not rejecting H_0 if H_1 is true is known as a **Type II Error** or a false dismissal.

The probability of a false alarm occurring is written α and known as the **significance** of the test. If H_0 uniquely determines a sampling distribution $P(\mathbf{x}|H_0)$, it is known as a **point hypothesis** and we can just write the probability that the data will end up in the critical region, assuming H_0 is true. For instance, in the example considered above, if n is large, the Central Limit Theorem tells us that the test statistic Z is a standard normal random variable, and therefore

$$\alpha = P(Z > 1.645|\mu = 0) \approx P(Z > z_{.95}|\mu = 0) \approx 0.05 \tag{2.13}$$

We say that the test has significance $\alpha = 5\%$. (Note this name is somewhat misleading, since rejecting H_0 with test with a smaller α level, would actually be a *more* significant result.) If the null hypothesis H_0 does not completely specify the sampling distribution (e.g., if we had chosen $\mu \leq 0$ rather than $\mu = 0$) it is known as a **composite hypothesis**, and α is defined to be the maximum of all the false alarm probabilities associated with the different distributions allowed by H_0. (In this case that would turn out to be when $\mu = 0$ anyway, so the significance would still be 5%.)

2.4.2 Power of a Test

The probability of a type II error (false dismissal) occurring, i.e., failing to reject the null hypothesis H_0 when the alternative hypothesis H_1 is true, is written β. The probability of *rejecting* H_0 when H_1 is true is called the **power** of the test, $\gamma = 1 - \beta$. Since the alternative hypothesis is often a composite hypothesis, the power of a test can depend on the value of any parameters that are not completely specified by H_1. We can talk about a **power curve** $\gamma(\theta)$ where θ is the parameter in question.

For example, in the test above based on the sample mean, the Central Limit Theorem still tells us that

$$\frac{\overline{X}-\mu}{s/\sqrt{n}} = Z - \frac{\mu}{\sigma/\sqrt{n}} \tag{2.14}$$

is standard-normal distributed whatever the value of μ, which means Z is normally distributed with unit variance but nonzero mean

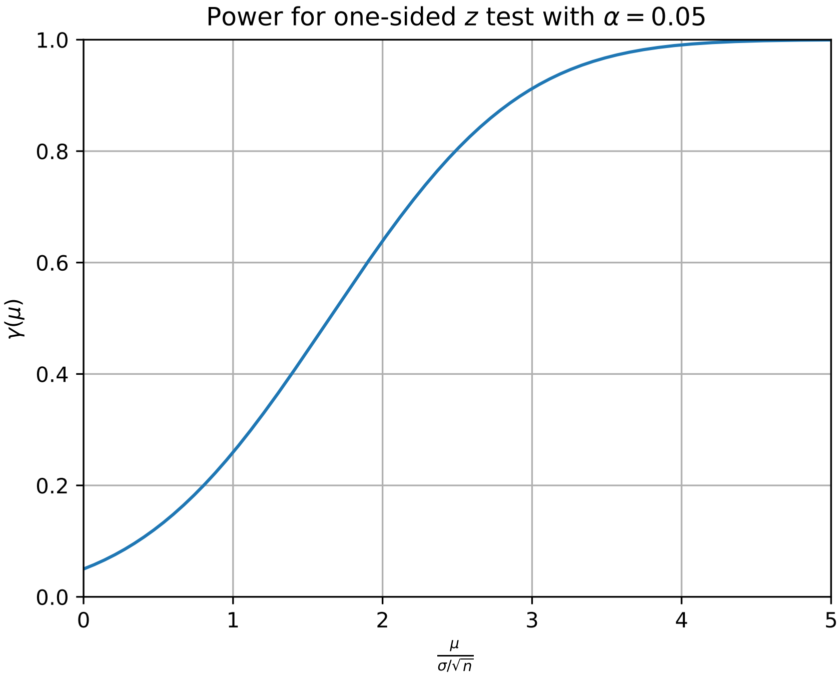
$$E(Z) = \frac{\mu}{\sigma/\sqrt{n}} \tag{2.15}$$

which means the power curve is

$$\gamma(\mu) = P(Z > 1.645|\mu) = 1 - \Phi\left(1.645 - \frac{\mu}{\sigma/\sqrt{n}}\right) \tag{2.16}$$

We can plot this using `ipython --pylab`:

```
from __future__ import division
import numpy as np
from scipy import stats
muscaled = np.linspace(0,5,1000)
power = stats.norm.sf(1.645-muscaled)
plot(muscaled,power)
title(r'Power for one-sided $z$ test with $\alpha=0.05$')
xlabel(r'$\frac{\mu}{\sigma/\sqrt{n}}$')
ylabel(r'$\gamma(\mu)$')
xlim(0,5)
ylim(0,1)
grid(True)
savefig('notes02_normpower.eps',bbox_inches='tight')
```



One thing to note is that the power $\gamma(\mu)$ goes to 0.05 (which is the significance α) as μ goes to 0 (which is the value specified by H_0). Also, $\gamma(\mu) > \alpha$ for all $\mu > 0$. This is a generally desirable property (the test should be more likely to reject H_0 when it's false than when it's true), and so it has a name. In general, an **unbiased test** is one for which $\gamma \geq \alpha$ for all possible point hypotheses contained within H_1.

Note that it quickly becomes impractical to calculate power curves analytically. Even in our example, if we'd used the statistic

$$T = \frac{\overline{X}}{\sqrt{S^2/n}} \tag{2.17}$$

constructed from the sample standard deviation s rather than the population standard deviation σ (which we might not know), it's not immediately obvious what to write down for the distribution of T when

$$T - \frac{\mu}{\sqrt{S^2/n}} \tag{2.18}$$

is approximately standard-normal distributed. We'll try to address these questions numerically whenever possible.

2.4.3 p-values

While these discussions of predefined tests with significance levels, rejection regions and power curves are easy to talk about in the abstract, they end up not providing so much information about an actual data set. Given such a test, the only result we have for an actual observed sample is "reject H_0" or "don't reject H_0". It is often more useful to consider how strongly H_0 can be rejected for a given observed sample $\{x_i\}$. This is the p-value, which is defined as the smallest significance level at which the null hypothesis would be rejected. It is also the probability, assuming the null hypothesis to be true, that a new sample drawn according to the same procedure, would give a test statistic at least as extreme as the one seen from the original data. (We can then also turn things around and say that a test with significance α tells us to calculate the p value, and reject H_0 if $p \leq \alpha$.)

Going back to our z-test example, we can compare H_0: $\mu = 0$ to H_1: $\mu > 0$ by calculating $z = \frac{\overline{x}}{\sigma/\sqrt{n}}$ and defining the p-value as

$$p = 1 - \Phi(z) \tag{2.19}$$

2.5 One-Tailed and Two-Tailed Tests

The example considered in detail so far included a one-sided alternative hypothesis H_1: $\mu > 0$, and so we rejected H_0: $\mu = 0$ if the value of z (and thus the value of $\overline{x}$) was too large (a one-tailed test). But we could also consider a two-sided alternative hypothesis H_1: $\mu \neq 0$ and reject H_0 if z is too large *or* too small. That's what's meant by "more extreme" in the p-value definition above. So for example, we could define a rejection region that says to reject H_0 if $z > 1.96 \approx z_{.975}$ or $z < -1.96$. This test will again have $\alpha = .05$. And if we instead want a p value from data giving us a statistic z, we'd have to ask for the probability that a standard normal random variable would be at least that far from 0:

$$p = \Phi(-|z|) + [1 - \Phi(|z|)] \tag{2.20}$$