ECE286 Abridged

Aman Bhargava

Janaury 2020

Contents

Chapter 1

Introduction

1.1 Counting

There are three main types of counting:

1. **With** replacement **with** ordering.
2. **No** replacement **with** ordering.
3. **No** replacement **No** ordering.

With replacement with ordering: The classic example here is counting the *number of possible passwords*. We have n options for each character (let's say $n = 26$) and we have sequences of k length (let's say the passwords are $k = 8$ long). Then our number of possible passwords is:

$$n^k$$

No replacement with ordering: You have n total distinct objects and you want to see how many k-long groups you can make with them. You can't use an object twice in the sequences. Your number of options goes down by 1 at every additional object added to a given sequences, so you end up with the following number of **permutations**:

$$(n)(n-1)(n-2)...(n-k+1) = \frac{n!}{(n-k)!} = nPr$$

No replacement no ordering: Same as above but the order doesn't matter (think of the number of **groups** instead of number of **sequences**). We already know that you can make nPr **sequences**, and that a given **group**

can be sequentialized in $k!$ ways. Therefore, we simply divide our answer for nPr by $k!$ as follows:

$$\frac{n!}{k!(n-k)!}$$

1.2 Random Experiments

Essentially experiments where you don't know the outcome in advanced and it is useful to think of them as having a *random* component. There are three hand-wavy types:

1. Designed: E.g. a coin toss.
2. Observational: Uncontrolled, e.g. observe and measure the time to get to school.
3. Retrospective: Looking at past data.

They all have **procedure** and **measurements**.

1.2.1 Sample Space

This is the set of all possible outcomes of the experiment, denoted by S. There are three types:

1. Finite (e.g. 3 coin tosses in a row)
2. Countably infinite (e.g. how many coin tosses until I get heads?)
3. Uncountably infinite (e.g. how tall is this person?)

1.2.2 Events

An event is a **set of outcomes** that we are interested in. For a coin flip experiment, we might have

$$A = \{HTT, THT, HHT\}$$

All events are **subsets** of the universal set S (sample space). The **Event Class** E is the set of all events. We assign probabilities p and relative frequences f to events in E.

$$\lim_{n\to\infty} f_A(n) = \lim_{n\to\infty} \frac{n_A}{n} = p_A$$

Chapter 2

Probability

2.1 Axioms of Probability Theory

Here's another list of three to remember:

1. For any event A, $P(A) \geq 0$.
2. $P(S) = 1$. Events will always come from the sample space S.
3. For any two disjoint sets $A \cap B = 0$ then $P(A \cup B) = P(A) + P(B)$.

One particularly nice property that arises from this is the following: for any events A and B,

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

You can get to this conclusion via the Venn diagram drawing approach.

2.2 Conditional Probability

The basic question we're answering here is: *How does event A affect the probability of B?*

Central Idea: because event A happened, the sample space shrinks to A for the next event.

Notation: The "probability of B given A" is denoted by $P(B|A)$.

$$P(B|A) \equiv \frac{P(A \cap B)}{P(A)}$$

Intuitively, this says that you divide the 'area' of the overlap region by the area of A to get $P(B|A)$.

2.2.1 Independence

A and B are independent if **knowing that B happened** doesn't tell you anything about A happening.

$$P(A|B) = P(A); \;\; P(B|A) = P(B)$$

It is easily demonstrable that it is a **symmetric property**. If B is independent of A, then A is independent of B.

There are two types of independence:

1. Contrived: Because of how we chose A and B - we kind of got lucky that it just happens to be.
2. The nature of the experiment: The two things are really fundamentally independent.

2.3 Total Probability

We define a partition through the following: The set $\{B_1, ..., B_n\}$ is a partition of S iff:

1. $B_i \cap B_j \; \forall i, j < k, i \neq j$
2. $\sum B_i = S$
3. For any event A:
$$A = \sum (A \cap B_i)$$

TOTAL PROBABILITY LAW:

$$P(A) = \sum_{i=1}^{k} P(A|B_i)P(B_i)$$

2.4 Bayes Theorem

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

A key idea is that the **partition** (in this case B) is usually the **INPUT**. We know $P(B)$ is known, $P(B|A)$ is measured in the experiment, then we find the $P(A|B)$. Think about the example where A is having cancer and B is the test for cancer being positive.

2.4.1 Multiple Tests

The classical question to test one's understanding of Bayes Theorem is one where a patient is tested for cancer. There is some probability $P(C)$ that the patient has cancer, there is a probability $P(\oplus)$ that the test will come back positive, and $P(\oplus|C)$. One is generally required to find $P(C|\oplus)$. With one test, it is a relatively straight forward plug-and play problem.

With two tests, things get a bit trickier. There are two keys to solving the problem:

- We assume that each test administered is **independent** of the other administered tests. This can be difficult to deeply understand due to the fact that it goes against our intuition of 'independent'. If we know that the first test resulted in a positive result, that changes the probability we would predict for the second to be positive. However, it is impossible to this problem without that assumption.
- The algebra for manipulating a conditional probability with multiple results is as follows:

$$P([\oplus\&\oplus]|C) = P(\oplus|C)P(\oplus|C)$$

Chapter 3

Random Variables and Probability Distributions

3.1 Random Variables

A random variable **maps** the results of an experiment to the real numbers. They **can be many** $\rightarrow$ **one** but **cannot** be **one** $\rightarrow$ **many**.

The range of all values that $X(s)$ (which is the random variable X) can take:

$$S_x = \{x | X(s) = x, s \in S\}$$

Generally speaking, big X is read as 'any random x value' or 'the set of possible x values' while little x is usually a given x value.

The difference between a discrete and continuous random variable is relatively straight forward. We have mapping functions f that assign probabilities to ouctomes x_i that obey the following properties:

1. $f(x_i) > 0$
2. $\sum f(x_i) = 1$
3. $P(x = x_i or x = x_j) = f(x_i) + f(x_j)$. This makes sense because one $\rightarrow$ many mappings are not allowed so the sets are definitely disjoint.
4. $\{s | X(s) = x_i\} \cap \{s | X(s) = x_j\} = 0$

3.2 Probability Functions

3.2.1 Probability Mass Functions (PMF's)

These are for discrete variables. We know that event $A = \{s|X(s) = x_i, s \in S\}$ and that $P(x = x_i) = P(A)$. The probability mass function is $f(x_i) = P(A)$ and has the following properties:

- $0 \leq f(x_i) \leq 1$
- $\sum f(x_i) = 1$
- If $A = \{x_1, x_2\}$ then $P(A) = f(x_1) + f(x_2)$.

3.2.2 Cumulative Distribution Functions (CDFs)

$$F(x) = P(X < x)$$

Read: F(x) is the probability that some random X is less than the given x.

Properties:

1. $0 \leq F(x) \leq 1$
2. $\lim_{x\to\infty} F(x) = 1$
3. $\lim_{x\to-\infty} F(x) = 0$
4. $F(x)$ is **non-decreasing**.

$$P(x < X < x + dx) = F(x + dx) - F(x)$$

3.2.3 Probability Density Functions (PDFs)

$$f(x) \equiv \lim_{dx\to 0} \frac{F(x + dx) - F(x)}{dx} = F'(x)$$

Properties:

1. $f(x) > 0$
2. $\int_{-\infty}^{\infty} f(x)dx = 1$
3. $P(a < X \leq b) = \int_a^b f(x)dx$

3.3 Mixed Random Variables

$$P(x < X < x + \Delta x) \approx f(x)\Delta x$$

3.4 Joint PDFs and PMFs

We are now switching to mapping outcomes to **vectors**. Each outcome s has mapping $[X(s), Y(s)]$.

Range: $S_{xy} - \{(x, y)|x = X(s), y = Y(s), s \in S\}$

3.4.1 Joint PMFs

$$f(x_i, y_j) = P(x = x_i, y = y_j)$$

Properties:

1. $0 \leq f(x_i, y_i) \leq 1$
2. $\sum\sum f(x_i, y_j) = 1$
3. If $A \subset S_{xy}$, $P(A) = \sum\sum_{(x_i,y_i)\in A} f(x_i, y_i)$

The general purpose here it to see the connection between two or more variables.

3.4.2 Marginal PMFs

$$g(x_i) = P(X = x_i) = \sum_{y_j \in S_y} f(x_i, y_i)$$
$$h(y_j) = P(Y = y_j) = \sum_{x_i \in S_x} f(x_i, y_i)$$

Interpretation:

3.4.3 Conditional PMFs

Conditional probability of y given $x = x_i$:

$$f(y_j|x_i) = P(Y = y_j|x = x_i) = \frac{f(x_i, y_j)}{g(x_i)}$$

Roughly the same rules apply to joint CDFs and PDFs, etc. Just replace the sums with integrals from negative to positive infinity.

Chapter 4

Mathematical Expectation

Means and variances. The main resource I use here is the textbook, so the notation may differ from the in-class notation. It is necessary to go back over the in-class material to correct the notation at a later date.

4.1 Expected Value

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

$$\mu = E[X] = \int_{-\infty}^{\infty} xf(x)dx$$

4.2 Variance

$$\sigma^2 = E[(X-\mu)^2] = \sum_x (x-\mu)^2 f(x)$$

$$\sigma^2 = E[(X-\mu)^2] = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx$$

Where σ^2 is the **variance** and σ is the **standard deviation**. μ is still the average or expected value. By making some simplifications, we also get:

$$\sigma^2 = E(X^2) - \mu^2$$

4.2.1 Covariance

For joint probability density/distribution function $f(x, y)$, the covariance of X and Y is:

$$\sigma_{XY} = E[(X - \mu_X)(Y - \mu_Y)] = \sum_X \sum_Y (x - \mu_X)(y - \mu_Y) f(x, y)$$

$$\sigma_{XY} = \iint (x - \mu_X)(y - \mu_Y) f(x, y)\, dx\, dy$$

This measures the **association** between the two. In other words, the amount of linear correlation. You can simplify it like before to be:

$$\sigma_{XY} = E(XY) - \mu_X \mu_Y$$

4.2.2 Correlation Coefficient

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

Properties:

- $-1 < \rho_{XY} < 1$
- Exact linear independency $\rightarrow \rho_{XY} = 1$ if $b > 0$ in $Y = mX + b$ and $\rho_{XY} = -1$ otherwise.

4.3 Means and Variance of Linear Combinations of Random Variables

Expected value $E()$ is LINEAR:

- $E(aX + b) = aE(X) + b$
- $E[g(X) \pm h(X)] = E(g(X)) \pm E(h(X))$. This also works for functions of two variables.
- $E(XY) = E(X)E(Y)$
- If X and Y are independent, $\sigma_{XY} = 0$

Theorem: $f(x, y)$ is a joint probability distribution and $a, b, c \in R$:

$$\sigma^2_{aX+bY+c} = a^2 \sigma^2_X b^2 \sigma^2_Y + 2ab\sigma_{XY}$$

Because $\sigma^2_{aX+bY+c} = E\{[(aX + bY + c)]^2 - \mu_{aX+bY+c}\}$. You then use linearity of all the operators to get to the final answer.

4.3.1 What if the Function is Non-Linear?

$$E(Z) = E(X/Y) \neq E(X)/E(Y)$$

For navigating non-linear functions of a random variable, we can take the taylor series expansion and **truncate** at the first linear term.

Chapter 5

Some Discrete Probability Distributions

5.1 Binomial and Multinomial Distribution

Let say that you have a random binary experiment that you run n times. Each trial can either be **successful** or **unsuccessful**. Binomial distributions tell you the probability of having k out of the n be **successful** given a probability p of each individual trail being successful.

5.1.1 Bernoulli Process

- Consists of **repeated trials**.
- Each trial has one of a **binary outcome**.
- Probability of success p remains the same from one trial to the next.
- Repeated trials are independent.

Binomial Distribution: Probability of distribution of a discrete random variable in a Bernoulli trials. If we let x be our random variable (representing the number of **successful** trials), the distribution of our probability mass function (PMF) is given by $b(x; n, p)$:

- Denoted by $b(x; n, p)$: depends on **number of trials and probability of success**.
- **Mean** and **variance** of $b(x; n, p)$ are: $\mu = np$, $\sigma^2 = npq$ respectively.

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

5.2 Multinomial Experiments and Distribution

What if our experiment has more than two potential outcomes? This is how multinomial distributions arise. As arguments, it takes the following (for a set of experiments with k possible outcomes and n trials):

- n: The number of trials run.
- $\vec{x}$: The hypothesized number of times each of the k outcomes will occur. $\sum x_i = n$, obviously.
- $\vec{p}$: The probability of each of the k possible outcomes.

The requirements of the trials themselves are very similar to those of a Bernoulli Process (independent outcomes, etc)

$$f(\vec{x}; \vec{p}, n) = \binom{n}{x_1, ..., x_k} p_1^{x_1} p_2^{x_2} ... p_k^{x_k}$$

where $\binom{n}{x_1,...,x_k} = \frac{n!}{x_1!...x_k!}$
Under the stipulation that $\sum \vec{x}_i = n$ and $\sum \vec{p}_i = 1$.

5.3 Hypergeometric Distribution

Basically the same as binomial distribution but without **statistical independence** between trials. Sampling at each trial is done **without replacement**.

It's exactly like the example of a deck of cards. We let the red cards be *successes* and the black cards be *failures*. We want to know the probability of getting a certain number of *successes* given n tries, **without replacing any of the cards**.

Utility: Binomial distributions are useful when you are looking to gauge the overall quality of a batch with *good* and *bad* labels. Hypergeometric distributions are there for when the test for good/bad is **destructive**. When that is the case, you can't replace the item to the sample bag, so independence is not maintained.

Formula

$$h(x; N, n, k) = \frac{\binom{k}{x}\binom{N-k}{n-x}}{\binom{N}{n}}$$

Where

1. n is the number of trials.
2. N is the total number of items being selected from at each trial.
3. k is the number of **successful** items available to be picked.
4. x is the *requested* number of successes (h gives the probability of selecting x successes by the end of the experiment). Under the stipulation

$$\max\{0, n - (N - k)\} \leq x \leq min\{n, k\}$$

5.3.1 Mean and Variance of Hypergeometric Distributions

$$\mu = \frac{nk}{N}$$

$$\sigma^2 = \frac{N-n}{N-1} n \frac{k}{N}(1 - \frac{k}{N})$$

When to approximate hypergeometric to binomial: If $n << N$ (the number of trials vs. the number of total items), then we can approximate a hypergeometric distribution as a binomial one. Specifically, if $n/N \leq 0.05$.

5.4 Negative Binomial Distribution

What if we begin a similar trial to before (counting the number of **successful** outcomes from random experiments), but instead of counting successes, we count the **number of trials it takes** to get a certain number of successful outcomes?

We are now interested in the probability of the kth success occurring on the xth trial. This is a **negative binomial** experiment.

Negative Binomial Random Variable: The number of trials (X) required to get k successes is the **negative binomial random variable**. Denoted by:

$$b^*(x; k, p) = \binom{x-1}{k-1} p^k q^{x-k}$$

b^* takes in the following and outputs the probability of it occurring:

- x: The number of trials taken to produce the result.
- k: The number of successes we are looking for.
- p: Probability of success at any given trial.
- Output b^*: Chance that it will take x trials to get k successes given probability p of success on a given trial.

5.5 Geometric Distribution

We have a standard binary trial procedure. A **geometric distribution** gives the chance that it will take x trials for the **first success** to occur:

$$g(x; p) = pq^{x-1}$$

Where $q = 1 - p$, the probability of failure at a given trial.

Mean and variance of Geometric Distribution:

$$\mu = \frac{1}{p}$$

$$\sigma^2 = \frac{1-p}{p^2}$$

5.6 Poisson Distribution and Poisson Process

Poisson Experiments: An experiment that gives the **number of outcomes in a time period or region** X.

Example: X is the number of telephone calls received per hour by an office.

Properties:

- The number of outcomes in a given interval is **independent** of the number of outcomes in a *disjoint* interval.
- Probability of an outcome occurring in a small interval is proportional to the length of that interval.
- The probability of more than one outcome occurring in a very small interval is negligible.

Formulae

$$\mu = \lambda t$$

Where μ is the average number of occurrences in spatial/temporal region t. λ is the proportionality between the two.

$$p(x; \lambda t) = \frac{e^{-\lambda t}(\lambda t)^x}{x!}, \;\; x = 0, 1, 2, ...$$

The Poisson probability sums give the chance that the number of occurrences is **less than or equal** to a value r over region t:

$$P(r; \lambda t) = \sum_{x=0}^{r} p(x; \lambda t)$$

Often found from tables.

Means and Variances:

$$\mu = \sigma^2 = \lambda t$$

Nature of Poisson Distributions:

1. More symmetric as average value grows large.
2. $b(x; n, p) \to p(x; \mu)$ as $n \to \infty$, $np \to \mu$ (p must go to zero for this to be true as well).

Chapter 6

Continuous Probability Functions

6.1 Continuous Uniform Distribution

This is just a flat distribution – anything in the given range is equally likely. Density function:

$$f(x; A, B) = \frac{1}{B-A} \quad \text{if } A \leq x \leq B, \;\; 0 \;\; \text{else}$$

Note that the value of $f(x)$ must integrate to 1 over its bounds.

$$\mu = \frac{A+B}{2}; \quad \sigma^2 = \frac{(B-A)^2}{12}$$

6.2 Normal Distribution

Also known as **Gaussian distribution**. One of the most important distributions in statistics, has extensive use in science and industry.

Normal random variable: A random variable X that is distributed according to a normal distribution. The probability density function is given as:

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

Importantly, as soon as μ, σ are specified, the entire distribution (density function) is parameterized.

Properties of a Normal Distribution:

- **Mode** occurs at $x = \mu$ (point where the density function is maximized).
- Curve is symmetric about μ.
- Inflection points occur at $x = \mu \pm \sigma$. Concave downward between the two inflection points, upward outside of them.
- Asymptotically approaches 0 as you get further from μ.
- Total area under the curve is 1.

6.3 Areas under the Normal Curve

There isn't actually a convenient form for the integral of the normal distribution function. We use tables to calculate it (or software).

For effiency: We can convert a normal random variable so that it has $mu = 0;\ \sigma = \sigma^2 = 1$. We call this adjusted variable Z where

$$Z = \frac{X - \mu}{\sigma}$$

We can now use tables for Z and use algebra to solve for the corresponding values of X. If we want the probability of $x_1 < X < x_2$, then we use the table to determine $z_1 < Z < z_2$ with $z_1 = (x_1 - \mu)/\sigma$, etc.

$$P(x_1 < X < x_2) = \frac{1}{\sqrt{2\pi}\sigma} \int_{x_1}^{x_2} e^{\frac{-1}{2\sigma^2}(x-\mu)^2} dx = \frac{1}{\sqrt{2\pi}} \int_{z_1}^{z_2} e^{-\frac{1}{2}z^2} dz$$
$$= \int_{z_1}^{z_2} n(z; 0, 1) dz = P(z_1 < Z < z_2)$$

STANDARD NORMAL DISTRIBUTION: When $\mu = 0;\ \sigma^2 = 1$ for a normal distribution.

Tables for making calculation more convenient often just tell you $P(Z < z)$.

6.4 Applications of the Normal Distribution

This section is relatively straight forward if one has a good grasp of how to use the normal distribution function.

6.5 Normal Approximation to the Binomial

As one might imagine, the line between discrete and continuous probability distributions can get blurred as we increase the number of elements in a binomial distribution to approach infinity.

Theorem: If X is a binomial random variable with $\mu = np$ and variance $\sigma^2 = npq$, then the *limiting form* of the distribution

$$Z = \frac{X - np}{\sqrt{npq}}$$

as $n \to \infty$ is the standard normal distribution $n(z; 0, 1)$.

Requirements for limit:

- $\mu = np$
- $\sigma^2 = npq$
- $n \to \infty$
- p, q aren't too close to 0 or 1.

Pretty good approximation even if n is small as long as $p \approx q \approx \frac{1}{2}$.

Continuity correction: If we want to know the probability of $X < x$ for a binomial distribution, we should take the integral from $(-\infty, x + 0.5]$ of the normal distribution. The $+0.5$ is called a 'continuity correction'.

6.6 Gamma and Exponential Distributions

The normal distribution is not universally perfectly applicable. The exponential distribution is just a special case of the gamma distribution, though, so they're in the same section of the textbook.

Important for *queuing theory* and *reliability probability*. Time between arrivals, time to failure of electrical parts, etc. are well modelled by exponential distributions.

6.6.1 Gamma Function

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx, \quad \text{for } \alpha > 0$$

Properties of the Gamma Function:

- $\Gamma(n) = (n-1)(n-2)(n-3)...(1)\Gamma(1)$ for any positive integer n.
- $\Gamma(n) = (n-1)!$ for a positive integer n.
- $\Gamma(1) = 1$.
- $\Gamma(\frac{1}{2}) = \sqrt{\pi}$

THE GAMMA DISTRIBUTION FUNCTION:

$$f(x;\alpha,\beta) = \frac{1}{\beta^{\alpha}\Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta}$$

when $x > 0$. Elsewhere, $f(x;\alpha,\beta) = 0$. Also, $\alpha, \beta > 0$.

6.6.2 Exponential Distribution

If we set $\alpha = 0$ in the gamma function, we get the exponential function.

$$f(x;\beta) = \frac{1}{\beta} e^{-x/\beta}$$

for $x > 0$ and $f(x;\beta) = 0$ elsewhere.

6.6.3 Means and Variances

$$\mu = \alpha\beta; \quad \sigma^2 = \alpha\beta^2$$

6.6.4 Relationship to Poisson Process

Poisson distribution if is for **counting discrete events** in a given **time period** or region. One can think of the period between events as a random variable in its own right (ex. time between arrivals at an airport).

Connection between Poisson distribution and exponential distribution comes when we look at the **probability of no events in a timeframe**:

$$p(0;\lambda t) = \frac{e^{-\lambda t}(\lambda t)^0}{0!} = e^{-\lambda t}$$

If we let X be the time before the first Poisson event, then the probability $P(X > x) = e^{-\lambda x}$. The CDF is therefore:

$$P(0 \leq X \leq x) = 1 - e^{-\lambda x}$$

Making the PDF

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

Which is **exactly the same** as an exponential distribution with $\lambda = 1/\beta$.

Applications of gamma/exponential distributions are therefore often to do with predicting the time-of-arrival for Poisson events.

6.6.5 Memoryless Property

Take the example of a component with an exponentially distributed lifetime prediction function. If we know that it has lasted to t_0 already, when we can say that

$$P(X \geq t) = P(X \geq t_0 + t | X \geq t_0)$$

Basically, if the piece has lasted for t_0 hours already, it has the same probability of lasting an additional t hours as it did at the beginning. The piece has no 'memory' of any damage it might have taken before.

However, if there is wear involves, then you should use a **gamma** or **Weibull distribution**.

Exponential distributions are good at describing **time between events** or time for 1 poisson event to occur. Gamma is good for describing the time for **multiple poisson events** to occur.

6.7 Chi-Squared Distribution

If we let $\alpha = v/2$ and $\beta = 2$ in the **gamma distribution** (v is a positive integer representing **degrees of freedom**), we get the **chi-squared distribution**.

$$f(x; v) = \frac{1}{2^{v/2}\Gamma(v/2)} x^{v/2-1} e^{-x/2}$$

for $x > 0$, v is the integer number of degrees of freedom. Important distributino for statistical hypothesis testing and estimation.

Mean and variance:

$$\mu = v; \quad \sigma^2 = 2v$$

6.8 Weibull Distribution

Weibull distribution is a common and effective way to model the lifetime of parts and assemblies. It is parameterized by two positive numbers α, β:

$$f(x;\alpha,\beta) = \alpha\beta x^{\beta-1}e^{-\alpha x^{\beta}}$$

for $x > 0$. f is 0 elsewhere. Alpha and beta must be greater than 0.

Mean and variance:

$$\mu = \alpha^{-1/\beta}\Gamma(1+\frac{1}{\beta});\quad \sigma^2 = \alpha^{-2/\beta}\{\Gamma(1+\frac{2}{\beta}) - [\Gamma(1+\frac{1}{\beta})]^2\}$$

Cumulative Distribution Function for Weibull distribution:

$$F(x) = 1 - e^{-\alpha x^{\beta}},\ \ x \geq 0$$

6.8.1 Failure Rate for Weibull Distribution

It is useful to predict the **probability that a part will function properly for AT LEAST time** t. We call that the reliability function $R(t)$.

$$R(t) = P(T > t) = \int_t^{\infty} f(t)dt = 1 - F(t)$$

Therefore the probability that a component will fail in the interval $T \in [t, t+\Delta t]$ is:

$$\frac{F(t+\Delta t) - F(t)}{R(t)}$$

As we take the limit for $\Delta t \to 0$, we get the **failure rate** $Z(t)$:

$$Z(t) = \lim_{\Delta t \to 0} \frac{F(t+\Delta t) - F(t)}{\Delta t}\frac{1}{R(t)} = \frac{f(t)}{1-F(t)}$$

$$Z(t) = \alpha\beta t^{\beta-1}$$

Interpretation of Failure Rate Quantifies the rate of change of the conditional probability that the component lasts an additional Δt given that it has lasted for t time. Some important properties are as follows:

- $\beta = 1$: Forms an exponential distribution with no *memory*. The component does not get more or less likely to break thanks to having lasted for time t.
- $\beta > 1$: Causes $Z(t)$ to increase with t, so the component wears down over time.

- $\beta < 1$: Causes $Z(t)$ to decrease with t, so the component strengthens over time.

Chapter 7

Functions of Random Variables

7.1 Transformations on Variables

If you have a function u and a discreete random variable X, you might want to know what the properties of $Y = u(X)$ are based on u and X. It has to be **one-to-one**, though.

Theorem 1: If X is a **discrete** random variable with probability distribution $f(x)$ and $Y = u(X)$ is a one-to-one transformation such that $y = u(x)$, $x = w(y)$, then the probability distribution function of y $g(y)$ is:

$$g(y) = f[w(y)]$$

Theorem 2: X_1, X_2 are **discrete** random variables with joint probability distribution $f(x_1, x_2)$. $Y_1 = u_1(X_1, X_2)$ and $Y_2 = u_2(X_1, X_2)$ are both one-to-one transformations between ordered pairs $(x_1, x_2), (y_1, y_2)$. We let $y_1 = u_1(x_1, x_2)$; $y_2 = u_2(x_1, x_2)$. Then the joint probability of y_1, y_2 is:

$$g(y_1, y_2) = f[w_1(y_1, y_2), w_2(y_1, y_2)]$$

Theorem 3: X is a **continuous** random variable with PDF $f(x)$. $Y = u(X)$ is a 1-1 relationship. $y = u(x)$; $x = w(y)$. Then

$$g(y) = f[w(y)]|J|$$

where $J = w'(y)$, or the **Jacobian** of the transformation.

Theorem 4: X_1, X_2 **cts** random vars with $f(x_1, x_2)$. $Y_1 = u_1(X_1, X_2)$; $Y_2 = u_2(X_1, X_2)$ are both 1-1. Then the joint PDF for the y's is:

$$g(y_1, y_2) = f[w_1(y_1, y_2), w_2(y_1, y_2)]|J|$$

Where

$$J = \det |\frac{\partial x_1}{\partial y_1}, \frac{\partial x_1}{\partial y_2}; \frac{\partial x_2}{\partial y_1}, \frac{\partial x_2}{\partial y_2}|$$

Theorem 5: X is **cts** RV with $f(x)$. $Y = u(X)$, and is **NOT 1-1**. If interval on X can be divided into k mutually disjoint sets that each have proper inverse functions $x_n = w_n(y)$, then the PDF of y is:

$$g(y) = \sum_{i=1}^{k} f[w_i(y)]|J_i|$$

Where $J_i = w_i'(y)$.

7.2 Moments and Moment-Generating Functions

Definition of Moment: The rth moment about the origin for RV X is:

$$\mu_r' = E(X^r) = \sum_x x^r f(x) = \int_{-\infty}^{\infty} x^r f(x) dx$$

Connections to known statistical properties include:

$$\mu = \mu_1'; \quad \sigma^2 = \mu_2' - \mu^2$$

7.2.1 Moment Generating Function

These form an alternatie function to determine the moments of a RV:

Definition: A **moment generating function** of X is given by $M_X(t)e^{tX}$. Therefore

$$M_X(t) = E(e^{tx}) = \sum_x e^{tx} f(x) = \int_{-\infty}^{\infty} e^{tx} f(x) dx$$

These exist only when the integral/sum converges. To get the actual moments from a moment-generating function, we do the following:

$$\mu_r' = \frac{d^r M_X(t)}{dt^r}|_{t=0}$$

The textbook has several good examples of solving things like binomial functions with this technique.

Uniqueness Theorem: If $M_X(t) = M_Y(t)$ for RV's X, Y for all t, then X, Y have the exact same probability distribution.

Addition Theorem:

$$M_{X+a}(t) = e^{at} M_X(t)$$

Multiplication Theorem:

$$M_{aX}(t) = M_X(at)$$

Sum Theorem: If $\{X_1, X_2, ..., X_n\}$ are independent random variables and $Y = \sum X_n$, then

$$M_Y(t) = M_{X_1}(t) M_{X_2}(t) ... M_{X_n}(t)$$

7.2.2 Linear Combinations of Random Variables

We let $Y = a_1 X_1 + a_2 X_2$ where both X are nomrally distributed with each having a μ, σ. We first find that

$$M_Y(t) = M_{X_1}(a_1 t) M_{X_2}(a_2 t)$$
$$M_Y(t) = \exp[(a_1\mu_1 + a_2\mu_2)t + (a_1^2\sigma_1^2 + a_2^2\sigma_2^2)t^2/2]$$

Hence the mean is $\mu = a_1\mu_1 + a_2\mu_2$, variance is $\sigma = a_1^2\sigma_1^2 + a_2^2\sigma_2^2$.

Theorem: Expanding for n summed normal distributions (independent), we get:

$$\mu_Y = \sum a_i \mu_i$$
$$\sigma_Y^2 = \sum a_i^2 \sigma_i^2$$

Chi Square Addition Theorem: If X_n each are mutually independent RV's with chi square distributions with v_n degrees of freedom respectively, then

$$Y = \sum X_n$$

Is yet another chi-squared distribution with $v = \sum v_n$ degrees of freedom.

Corrolary: Normal Sum to Chi-Square: Each X_n are independent RV's with identical normal distributions (same mu, σ). Then

$$Y = \sum_{i=1}^{n} (\frac{X_i - \mu}{\sigma})^2$$

has a chi-squared distribution with $v = n$.

Corrolary: Expanded Normal Sum to Chi-Squared: Now each of the X_n RV's can have different means and standard deviations. Now

$$Y = \sum_{i=1}^{n} (\frac{X_i - \mu_i}{\sigma_i})^2$$

has a chi-squared distribution with $v = n$.

Chapter 8

Introductory Statistics and Data Analysis Topics

8.1 Overview: Statistical Inference and Probability

Vocabulary

- **Inferential statistics**: A toolbox for making scientific judgements in the face of variablility and uncertainty.
- **Sources of variation**
- **Samples**: Collections of *observations*.
- Two main types of studies: observational and controlled.
- **Descriptive statistics**: Used when you want a summary of a dataset. Measures of central tendency, variation, etc.

P-Values Let's say that a process yields 10 defective components out of 100 sampled ones. If we say that the **maximum acceptable error rate** is 5%, we can calculate that the probability that 10 or more out of 100 were defective would be 0.002 if the true error rate was 5%. That is our P-value for an error rate of 5%. From this, we learn that the probability that we are actually OK in terms of true error rate is incredibly small.

P-values can also be used when one wants to know whether or not there is a statistical difference between observations on two population. It can measure **the probability that these results would be obtained if there were no true difference**.

Probability vs. Statistical Inference Inferential statistics uses probability theory to draw conclusions about a dataset.

Probability theory lets you draw conclusions about hypothetical data that you know some features about in a deductive fashion.

You are taught probability theory first to substantiate the algorithms in statistical theory.

8.2 Sampling Procedures and Data Collection

Simple Random Sampling Characterized by each sample within a set having **equal likelihood** of being sampled. Usually is the gold standard for mitigating bias, but it is sometimes advantageous to use other sampling methods. *Stratified random sampling* is used when the population isn't homogenous and consists of *strata* – non-overlapping groups. With stratified random sampling, you would perform random sampling on each stratum.

Experimental Design

- **Treatments**: Different groups in an experiment can be subjected to treaments or treatment combinations.
- **Experimental unit**: The different groups in an experiment.
- **Completely Randomized Design**: Participants are assigned to experimental groups entirely at random. Done to make sure that extraneous characteristics of each group do not overpower the treatment in question.

8.3 Measures of Location: Mean and Median

Definition: Sample Mean $\bar{x}$ is given by

$$\bar{x} = \sum_{i=1}^{n} x_i/n$$

Definition: Sample Median is given by the following assuming that $x_1, ..., x_n$ are arranged in INCREASING ORDER.

$$\tilde{x} = x_{(n+1)/2}; \quad \text{if } n \text{ is odd}$$

$$\tilde{x} = \frac{1}{2}(x_{n/2} + x_{n/2+1}) \quad \text{if } n \text{ is even}$$

8.4 Measures of Variability

Sample Variance: Represented by s^2 is given by:

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

Sample Standard Deviation: Denoted by s, given by:

$$s = \sqrt{s^2}$$

The $n-1$ is often called the **degrees of freedom associated with the variance** estimate.

8.5 Statistical Modeling, Scientific Inspection, and Graphical Diagnostics

The section on continuous vs. discrete data was foregone as it is a relatively simple distinction.

Postulated model: Often at the end of analysis, the parameters of a *postulated model* are revealed. We can also do graphical analysis:

- **Scatter plot:** If you don't know what a scatter plot is, you have some problems.
- **Stem-and-Leaf plot:** Left column is everything except for the ones place. The right column is the ones place for each sample.

Chapter 9

Fundamental Sampling Distributions and Data Descriptions

9.1 Random Sampling

Populations and Samples **Population:** The 'totality' of all the observations we are interested in. *Size* is determined by the number of observations in the population.

Sample: A subset of the population.

We make inferences from the samples to the populations. To ensure the validity of these inferences, we need to make sure that the sample selection protocol was **unbiased**.

Random Sampling: When observations are made independently and at random.

We let each X_i represent the ith sampling from the population. $X_1, ..., X_n$ constitute a *random sample* from the population with values $x_1, ..., x_n$. If $f(x)$ is the probability distribution function and all the measurements that comprise the random sample are independent, we can make a **joint probability distribution** function to determine the probability of a collecting a given random sample:

$$f(x_1, ..., x_n) = f(x_1)f(x_2)...f(x_n)$$

9.2 Important Statistics

Definition: A **statistic** is any function of the random variables that make up a random sample.

9.3 Sampling Distributions

Definition: The **sampling distribution** is the probability distribution of a *statistic*.

9.4 Sampling Distribution of Means and the Central Limit Theorem

Sampling Distribution of $\bar{X}$ We assume n samples were taken from a *normal population* with mean μ and variance σ^2. From our established theorems, we get:

$$\bar{X} = \frac{1}{n}\sum X_i$$

$$\mu_{\bar{X}} = \frac{1}{n}(\mu + \mu + \mu + ...) = \mu$$

$$\sigma^2_{\bar{X}} = \frac{1}{n^2}(\sigma^2 + \sigma^2 + ...) = \sigma^2/n$$

Central Limit Theorem: If $\bar{X}$ is the mean from n samples from a normal population with distribution mu, σ^2, then the limiting form of the distribution

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

as $n \to \infty$ is the standard normal $n(z; 0, 1)$. This is generally good for $n \geq 30$ if the distribution isn't too skewed. Perfect regardless of n if the sampled distribution is perfectly normally distributed.

9.4.1 Inferences on the Population Mean

The central limit theorem is very useful for inferring information about the true distribution of a stochastic process from n samplings. It's pretty clear how this would be applied – if you take a sample of size n and you determine

some average $\bar{x}$, you can take the integral of the **sampling distribution** for $\bar{X}$ with the lower or upper bound as $\bar{x}$ to determine the probability that the true mean is μ.

9.4.2 Sampling Distribution between Two Means

We want to compare two populations X_1, X_2 that have, respectively, $\bar{X}_1, \bar{X}_2, \sigma_1, \sigma_2, \mu_1, \mu_2$. We take a sample of size n_1, n_2 from each population, we want to find statistics about $\bar{X}_1 - \bar{X}_2$:

$$\mu_{\bar{X}_1 - \bar{X}_2} = \mu_{\bar{X}_1} - \mu_{\bar{X}_2} = \mu_1 - \mu_2$$
$$\sigma^2_{\bar{X}_1 - \bar{X}_2} = \sigma^2_{\bar{X}_1} + \sigma^2_{\bar{X}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

The following equation, therefore, approximates the standard normal Z:

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{(\sigma_1^2/n_1) + (\sigma_2^2/n_2)}}$$

As again, this approximation generally works best when $n \geq 30$.

9.5 Sampling Distribution of S^2

Notation: S^2 is the variance of a sample from a distribution. σ^2 is the true variance of the random variable.

Theorem: If we take a sample of size n from a population that is known to have variance and mean σ, μ and we get a measured variance of S^2 from our sample, then

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{\sigma^2}$$

is a chi-squared distribution with $v = n - 1$ degrees of freedom.

You can calculate the value of χ^2 for a given sample by $\chi^2 = \frac{(n-1)s^2}{\sigma^2}$. The chance that a χ^2 value greater than or equal to the one returned by that function can be found via some tables that integrate the chi-squared probability density function.

The common method of interpreting this is as follows: Since 95% of the chi-square function falls between $\chi^2_{0.025}$ and $\chi^2_{0.975}$. Your steps to solving a problem of this type are as follows:

1. Calculate the variance of the sample s^2.

2. Assuming that the initial distribution was normal, use the given value of σ^2 to calculate χ^2 using one of the above formulae:

3. Depending on the number of degrees of freedom $v = n-1$, determine if the calculated χ^2 falls in that 95% region. If it does, the current value of σ^2 is fine!

9.6 t-Distribution

The *central limit theorem* assumes that you know the true σ of the process you are dealing with. This isn't always the case, so t-distributions allow you to get roughly the same functionality as the central limit theorem by approximating the true σ with the s value of the sample.

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

A small sample size makes $S \neq \sigma$ very likely, so the t-distribution deviates substantially from the standard normal.

t-distribution Theorem: Z is the standard normal. V is a chi-squared random variable with v degrees of freedom. Z, V are independent. Then the random variable T is given by

$$T = \frac{z}{\sqrt{V/v}}$$

has the probability density function

$$h(t) = \frac{\Gamma[(v+1)/2]}{\Gamma(v/2)\sqrt{\pi v}}(1 + \frac{t^2}{v})^{-(v+1)/2}, \ \ -\infty < t < \infty$$

The utility of this is the following:

Utility of t-distribution: If $X_1, X_2, ..., X_n$ are random variables with a mystery shared μ, σ, and we lett

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

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Then we get

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

That has a t-distribution with $v = n - 1$ degrees of freedom.

Using the t-distribution:

- Basically looks the same as a normal distribution, is a bit wider because of greater variance in values.
- t_α represents the value of t above which there is an area of α.
- t-distributions are symmetrical about zero. Therefore, $t_{1-\alpha} = -t_\alpha$.
- 95% of the t-distribution lies between $-t_0.025$ and $t_0.025$.

9.7 F-distribution

t-distributions is really useful for comparing the *means* of two populations when you also don't know the variance of either. F-distributions are really useful for comparing the *variances* of two populations when you have limited information.

F-distribution Theorem: U and V are chi-squared distributions with v_1, v_2 degrees of freedom respectively. The probability distribution of random variable F

$$F = \frac{U/v_1}{V/v_2}$$

is

$$h(f) = \frac{\Gamma[(v_1 + v_2)/2](v_1/v_2)^{v_1/2}}{\Gamma(v_1/2)\Gamma(v_2/2)} \frac{f^{(v_1/2)-1}}{(1 + v_1 f/v_2^{(v_1+v_2)/2})}$$

for values of $f \geq 0$.

As per usual, f_α represents the value of f where the area above that value is equal to α. A convenient switching theorem is:

$$f_{1-\alpha}(v_1, v_2) = \frac{1}{f_\alpha(v_2, v_1)}$$

9.7.1 F-distribution with Two Sample Variances

Samples of size n_1, n_2 are selected from populations with σ_1^2, σ_2^2 respectively. From our established theorems on chi-squared relationships with randomly selected values from normal distributions,

$$\chi_1^2 = \frac{(n_1 - 1)S_1^2}{\sigma_1^2}; \quad \chi_2^2 = \frac{(n_2 - 1)S_2^2}{\sigma_2^2}$$

are chi-squared distributions with $v_1 = n_1 - 1, v_2 = n_2 - 1$ degrees of freedom respectively. We let $X_1^2 = U$ and $X_2^2 = V$. From there we get this very applicable result:

Comparing sample variances of independent random variables:

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} = \frac{\sigma_2^2 S_1^2}{\sigma_1^2 S_2^2}$$

has an F-distributtion with $v_1 = n_1 - 1$ and $v_2 = n_2 - 1$ degrees of freedom.

9.8 Quantile and Probability Plots

9.8.1 Quantile Plots

Definition: the quantile of a sample $q(f)$ is the value for which the fraction f of the provided data is less than. f goes from 0-1. $q(0.1)$ returns a number q. $P(F < q) = f$, you could think about it as.

9.8.2 Normal Quantile-Quantile Plot

The rest of this note has been excluded as the syllabus has been revised to only include 8.1-6.

Chapter 10

One and Two-Sample Estimation Problems

This chapter deals with statistical inference and estimating population parameters. Estimation procedures discussed will only involve one and two samples.

10.1 Statistical Inference

The classical approach to statistical inference is the **classical method** where we directly infer the parameters of a population based on the data we gather. The **Bayesian method** leverages prior knowledge about the population with gathered information to form conclusions about the data.

Estimation vs. hypothesis testing: two main classes of statistical inferences. Estimation is a regression problem: you are trying to approximate some ground truth value from the world. Hypothesis testing is a boolean classification problem where you are trying to determine the validity of a conjecture.

10.2 Classical Method of Estimation

Point estimate: A single actual value of a theoretical statistic. For the statistic $\bar{X}$ (the average), $\bar{x}$ is the point estimate. $\bar{x}$ is based off of some n readings and approximates some actual true fact about the distribution.

Unbiased estimator: If an estimator has a mean equal to the parameter being approximated, then it is **unbiased**.

$$\mu_{\hat{\Theta}} = E(\hat{\Theta} = \theta$$

Where $\hat{\Theta}$ is the estimator and θ is the true value.

Estimator efficiency: If two estimators approximate the same true population parameter θ and are **unbiased**, then the one with the lower variance is the **more efficient estimator.**

Interval Estimation: This is exactly what it sounds like. It is an estimation of the interval between which one is likely to find the true value of a population parameter.

$$\hat{\theta_L} < \theta < \hat{\theta_U}$$

The length of an interval estimation indicates the accuracy of the central point measurement.

Interpreting Interval Estimates: The actual values $\hat{\theta}_L, \hat{\theta}_U$ correspond to random variables $\hat{\Theta}_{L,U}$ such that

$$P(\hat{\Theta}_L < \theta < \hat{\Theta}_U) = 1 - \alpha$$

Where $100(1-\alpha)\%$ is the **confidence interval** and $1-\alpha$ is the **confidence coefficient/degree of confidence**. The lower and upper bounds are the **confidence limits**.

10.3 Single Sample: Estimating the Mean

- Sampling distribution of $\bar{X}$ has center μ and is generally the best estimator of μ.
- $\bar{x}$ is the **point estimate**.
- $\sigma^2_{\bar{X}} = \sigma^2/n$
- We can construct the **confidence interval** for our estimate of μ based on the sampling distribution of $\bar{X}$

Confidence Interval on μ if σ^2 is known: $\bar{x}$ is the mean of a random sample size n from a population with known variance σ^2. A confidence interval of $100(1-\alpha)\%$ for approximating μ isgiven by:

$$\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Where $z_{\alpha/2}$ is a z-value that leaves area $\alpha/2$ to the right.

Good results are guaranteed by this theory assuming $n \geq 30$ and the distribution is not very skewed.

A nice resultant theorem is as follows:

Theorem on Error of μ approximation: If $\bar{x}$ is used as an estimate of μ, our confidence is $100(1-\alpha)\%$ that the error of our estimate will not exceed $z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$.

Inverse Theorem on μ approximation: If we want to confine error to e with $100(1-\alpha)\%$ confidence for our estimate $\bar{x}$ of μ, our number of samples must be:

$$n = (\frac{z_{\alpha/2}\sigma}{e})^2$$

10.3.1 One-Sided Confidence Bounds

By the central limit theorem:

$$P(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < z_\alpha) = 1 - \alpha$$

Which can be manipulated to...

$$P(\mu < \bar{X} + z_\alpha \sigma/\sqrt{n}) = 1 - \alpha$$

10.3.2 Case of Unknown σ

Recall the T-distribution:

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

The random variable T has a t-distribution with $n-1$ degrees of freedom. S represents the sample standard deviation. Now we can use the T distribution

to construct our confidence intervals, replacing the normal distributions from before with T distributions:

$$P(-t_{\alpha/2} < T < t_{\alpha/2}) = 1 - \alpha$$

$$P(-t_{\alpha/2} < \frac{\bar{X} - \mu}{S/\sqrt{n}} < t_{\alpha/2}) = 1 - \alpha$$

$$P(\bar{X} - t_{\alpha/2}\frac{S}{\sqrt{n}} < \mu < \bar{X} + t_{\alpha/2}\frac{S}{\sqrt{n}}) = 1 - \alpha$$

Theorem on confidence interval on μ with σ^2 unknown: $\bar{x}, s$ are the sample mean and standard deviation. The $100(1-\alpha)\%$ confidence interval for μ is:

$$\bar{x} - t_{\alpha/2}\frac{s}{n} < \mu < \bar{x} + t_{\alpha/2}\frac{s}{\sqrt{n}}$$

where $t_{\alpha/2}$ is the value on the t distribution of $v = n-1$ degrees of freedom that has an area of $\alpha/2$ to the right.

Large-Sample Confidence Intervals: If $n \geq 30$, it is recommended that s can replace σ, so

$$\mu \approx \bar{x} \pm z_{\alpha/2}\frac{s}{\sqrt{n}}$$

This is known as the *large-sample confidence interval.*

10.4 Standard Error on a Point Estimate

We now equate the **standard deviation** of an estimator with the **standard error** of the estimator (notated as s.e.$(\hat{\theta})$. We write:

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = \bar{x} \pm z_{\alpha/2}\text{s.e.}(\bar{x})$$

10.5 Prediction Intervals

The intervals we have gone through so far are about our confidence that our **mean value** that we calculated based off of some observations is the correct one. This is appreciably different from when we have to figure out

our confidence that an individual component is defective or not as the latter depends heavily on the standard deviation of the population.

Prediction interval of Future Observation (σ is known): A $100(1-\alpha)\%$ **prediction interval** of a future observation x_0 for a population that has an *unknown mean* μ and *known variance* σ^2 is:

$$\bar{x} - z_{\alpha/2}\sigma\sqrt{1+1/n} < x_0 < \bar{x} + z_{\alpha/2}\sigma\sqrt{1+1/n}$$

Where $z_{\alpha/2}$ is the z-value leaving area $\alpha/2$ to the right.

Prediction interval of Future Observation, σ^2 is unknown:

$$\bar{x} - t_{\alpha/2}s\sqrt{1+1/n} < x_0 < \bar{x} + t_{\alpha/2}s\sqrt{1+1/n}$$

Where $t_{\alpha/2}$ is the t-value with $v = n-1$ degrees of freedom that leaves $\alpha/2$ to the right.

One-sided predictions: Upper bound is $\bar{x} + t_\alpha s\sqrt{1+1/n}$ and the lower bounded one-sided prediction would be $\bar{x} - t_\alpha s\sqrt{1+1/n}$.

10.5.1 Prediction Limits for Outlier Detection

Outlier detection rule: An observation is an outlier if it falls outside the prediction interval computed without including the questionable observation in the sample.

10.6 Cramer-Rao Lower Bounds

The goal here is to establish that, for any unbiased estimator ($E(\hat{\theta}) = \theta$), the variance of the estimator is always greater than some lower bound called the **Cramer-Rao lower bound**

CRLB Theorem: If the joint pdf satisfies the regularity condition

$$E(\frac{\partial}{\partial\theta}\ln f(x,\theta)) = 0$$

Then for **any** unbiased estimator we have the **lower bound**

$$\text{var}\hat{\theta} \geq \frac{-1}{E[\frac{\partial^2}{\partial\theta^2}lnf(x;\theta)]}$$

Chapter 11

One and Two-Sample Tests of Hypotheses

11.1 Statistical Hypotheses: General Concepts

Statistical Hypothesis: an assertion or conjecture concerning one or more populations.

Role of Probability in Hypothesis Testing

- The rejection of a hypothesis implies evidence that refutes it.
- Rejection implies that, if the if the hypothesis was true, there was a small probability of making the observation in question.
- **That being said,** failure to reject the hypothesis does not rule out any possibilities.
- Therefore, if you want to **strongly support** some conclusion, you should do so via the rejection of a hypothesis.

Null and Alternative Hypothesis **Null hypothesis:** any hypothesis H_0 we wish to test. Rejection results in the **acceptance of the alternative hypothesis** H_1, which is usually the actual question we want answered. Your two possibilities for the result of your experiment include:

1. **Reject** H_0 in favour of H_1 because of sufficient evidence.
2. **Fail to reject** H_0 because of insufficient evidence in data.

11.2 Testing a Statistical Hypothesis

Test Statistic: The number we measure in order to make our final decision about the hypothesis.

Critical region: Values of our test statistic that would result in the rejection of the null hypothesis.

Critical value: Boundary number between criticla region and non-critical region.

11.2.1 Probability of a Type I Error

Type I error: when you incorrectly reject the null hypothesis.

Type II error: when you fail to reject the null hypothesis even though the null hypothesis is actually false.

	H_0 **is true**	H_0 **is false**
Do not Reject H_0	Correct	Type II Error
Reject H_0	Type I Error	Correct

Level of Significance: The probability of a type I error α

In th example of testing whether a new vaccine is better than the old one that only works for a quarter of the people after two years:

$$\alpha = P(\text{type I error}) = P(X > 8 \,\text{when}\, p = \frac{1}{4}) = \sum_{x=9}^{20} b(x; 20, \frac{1}{4}) = 1 - 0.9591 = 0.0409$$

We say that *we are testing the null hypothesis at the* $\alpha = 0.0409$ *level of significance* (a.k.a. *size of test*). Since this value is very small, it is unlikely that a type I error will be committed.

Probability of a Type II Error: denoted by β, is **impossible to compute unless there is a specific laternative hypothesis**. In the case of the vaccine testing experiment, our alternative hypothesis would have to take on the form $p = p_0$ (let say $p = \frac{1}{2}$ for the example...)

$$\beta = P(\text{type II error}) = P(X \leq 8 \,\text{when} p = \frac{1}{2})$$

$$\beta = \sum_{x=0}^{8} b(x; 20, \frac{1}{2}) = 0.2517$$

This is a pretty large value because it is likely that, even if the vaccine works twice as well as the old one, we will risk accepting the null hypothesis. Ideally we obviously want low α, β

We can adjust α, β by changing our critical value. We can also decrease both simutlaneously by increasing our sample size.

Power: The power of a test is the *probability of rejecting* H_0 in the case that a specific alternative hypothesis is true, $= 1 - \beta$.

Example: We have a population and our null hypothesis is that the mean is NOT 68. That makes $H_1 : \mu \neq 68$. Ourvalue for β when our alternative hypothesis is set to $\mu = 68.5$ is $\beta = 0.8661$, making the **power** $1 - 0.8661 = 0.1339$. We then say that **our test will properly reject** H_0 **only 13.39% of the time if the true mean is 68.5**.

11.2.2 One- and Two-Tailed Tests

If the hypotheses are of the form

$$H_0 : \ \theta = \theta_0; \ \ H_1 : \ \theta > \theta_0$$

then it is a **one-tailed** test. The critical region is in one tail of the distribution, whether it is the right tail or left tail.

If the hypotheses are of the form

$$H_0 : \ \theta = \theta_0 \ \ H_1 : \ \theta \neq \theta_0$$

then it is a **two-tailed** test for similar reasons as described above.

Choosing a Null and Alternative Hypothesis

- H_0 often has an *equals sign.*
- Make sure that you select the right number of tails – if it is important to detect when something is both super or inferior, use two-tailed.

11.3 P-Values for Decision making in Testing Hypotheses

It is customary to use α values of 0.05 or 0.01 to select the critical region. If we go with $\alpha = 0.05$ and we have a two-tailed test involving the standard normal, then our critical region would be

$$z > 1.96 \text{ or } z < -1.96$$

Because $z_{0.025} = 1.96$ and we need to split the $\alpha = 0.05$ across both tails.

The dogmatic obsession with P-values of $\alpha = 0.05$ doens't always make sense – if you need to adjust your P-value to 0.06, you aren't really increasing your probability of committing a type I error that much. It's good practice to show calculate the P-value that would be required to reject the null hypothesis no matter what your initial conclusion is with a conventional P-value of 0.01 or 0.05.

P-Value Definition: P is the lowest possible value of significance α at which the observed value of the test statistic is considered significant.

The P-value approach is considered more modern than the fixed-α approach. Here are the procedural steps to follow for classical and P-value approach.

Classical Fixed-α

- State null and alternative hypotheses.
- Choose a fixed level of signficance α.
- Choose an appropriate test statistic to establish critical region on α
- Reject H_0 if the test statistic is in the critical region. Otherwise, do not reject.
- Draw conclusions.

P-value approach

- State null and alternative hypothesis.
- Select test statistic.
- Compute the P-value based on computed value of test statistic.
- Use your judgement of the P-value to draw conclusion.

11.4 Not-In-Textbook

- **Q-Function**: Let Z be standard normal. We define the Q-function as $Q(z) = P(Z > z) = \int_z^\infty \frac{1}{\sqrt{2\pi}} e^{t^2/2} dt$.

Chapter 12

Logistics

12.1 Midterm I

- Date: Thursday, March 5.
- Covers chapters 2-5.