Lecture Notes for
Statistical Review: Prerequisites

Random Variables

A random variable associates a numerical value with each outcome of an experiment. A random variable is a function from the sample space to real numbers. In notation:

$$X : \Omega \to \mathbb{R} \tag{1}$$

Remember that Ω is the sample space of an experiment. Remember that a function, say $f : \mathcal{A} \to \mathcal{B}$ associates an element of $\mathcal{A}$ with an element of $\mathcal{B}$.

Discrete Random variable

We call a random variable *discrete* if it has a countable number of values. A list, table, or functional specification of the numerical values of X, a random variable, and the probabilities associated with it is the *probability distribution of a random variable*.

It often helps to write the probability X equals x, $\mathbb{P}(X = x)$ as $p(x)$. In this way we functionalize the probability statement. I read $\mathbb{P}(X = 2) = p(2)$ as the probability X equals two.[1]

The following equation is a fact by extension of the probability rule to random variables:

$$\sum_{all\ possible\ value\ of\ X} p(x) = \sum_{all\ possible\ value\ of\ X} \mathbb{P}(X = x) = 1 \tag{2}$$

The *cumulative distribution function* is

$$F(x) = \sum_{all\ t \leq x} p(t) \tag{3}$$

or $F(x) = \mathbb{P}\{X \leq x\}$.

The *mean*, μ, and variance, σ_2 are *population parameters* which correspond to the sample. We weight each value by its probability p(x); the quantities are, in a sense, weighted averages. The mean or *expected value* of X is

$$\mu = \sum_{all\ X} xp(x) = \sum_{all\ X} x\mathbb{P}(X = x) = \mathbf{E}(X) \tag{4}$$

and the variance is

$$\sigma^2 = \mathbf{E}[(X - \mu)^2] = \sum_{all\ X} (x - \mu)^2 p(x) = Var(X) \tag{5}$$

The *Standard Deviation* of X is just the square root of the Variance $\sqrt{\sigma_X^2} = \sigma$.

Means and Variances of Linear Transforms. A *linear transform* is any $aX + b$ of X. If X has mean μ_X, and a, b are constants (known), then:

$$\mathbf{E}(aX+b) = \sum (ax+b)p(x) = \sum a{\cdot}x{\cdot}p(x) + \sum b{\cdot}p(x) = a\sum x{\cdot}p(x) + b\sum p(x) = a\mathbf{E}(X)+b = a\mu_X+b \tag{6}$$

[1] Notice the capital X vs. the lower case x. The first is the random variable and the second is its observed value

If Y has mean μ_y, and c, d are constants (known) as well, then a natural extension is:

$$\mathbf{E}(aX + b + cY + d) = a\mu_X + c\mu_y + b + d \tag{7}$$

Variance of the linear transform is

$$Var(X) = \sigma^2 = \mathbf{E}[(X - \mu)^2]$$

We can write $\mathbf{E}[(X - \mu)^2] = \mathbf{E}[X^2 - 2X\mu + \mu^2] = \mathbf{E}(X^2) - 2\mu\mathbf{E}(X) + \mu^2$. Which means

$$\sigma^2 = \mathbf{E}(X^2) - 2\mu\mathbf{E}(X) + \mu^2$$

and thus

$$Var(X) = \mathbf{E}(X^2) - [\mathbf{E}(X)]^2 \tag{8}$$

In words, the *variance is the difference between the squared expectation and the expectation squared*. You should remember this.

Now the variance for the general linear transform: aX + b.

$$Var(aX + b) = a^2(\mathbf{E}(X^2) - \mu_X^2) = a^2 Var(X) \tag{9}$$

Since

$$\mathbf{E}[(aX + b)^2] - [\mathbf{E}[(aX + b)]]^2 = a^2\mathbf{E}(X^2) + 2ab\mu_X + b^2 - a^2\mu_X - 2ab\mu_X - b^2$$

.

Note that *adding a constant does not increase the variance*.

As it turns out, the expectation of any function of X is just

$$E[g(X)] = \sum g(x)p(x) \tag{10}$$

This is <u>not</u> true for the variance.

Continuous Random Variables

We call a random variable *continuous* if it has an uncountable number of values; if it can take all values in an interval of values. For X a continuous random variable, the *pdf* of X is the function $f(X)$ such that:

$$\mathbb{P}(a \leq X \leq b) = \int_a^b f(x)dx. \tag{11}$$

We call, $f(x)$, the density curve for X. We can also restate some of our probability rules using this new definition. For X a continuous random variable, on the real line, with density function $f(x)$

- $\int_{-\infty}^{x} f(u)du = F(x)$. $F(x)$ is called the distribution function for X. $F(x) = \mathbb{P}(X \leq x)$, or the probability that and random variable X is less than or equal to x.

- $\int_{-\infty}^{+\infty} f(x)dx = F(+\infty) - F(-\infty) = 1 - 0 = 1$. Pay attention to nuance here: The distribution function of X is 1 at infinity. Every value of X is less than or equal to infinity. There is an analogous argument for $F(-\infty) = 0$. And I point out that the area under the density curve **must** equal 1.

- For all $X \in]-\infty, +\infty[$, $0 \leq f(x) \leq 1$

The property of the complement yields:

$$\overline{F}(x) \equiv \mathbb{P}(X > x) = 1 - \mathbb{P}(X \leq x) = 1 - F(x) \tag{12}$$

$1 - F(x)$ is called the *survival distribution* for X.
The *median* $\widetilde{\mu}$ for the discrete distribution is

$$\widetilde{\mu} = \{x s.t. F(x) = 0.5\} \tag{13}$$

For a continuous R.V. the *median* is

$$\widetilde{\mu} = \{x \; s.t. \int_{-\infty}^{x} f(u)du = .5\} \tag{14}$$

The median is a special case of a *p-tile* or *percentile* of the distribution of X. In general, the pth p-tile, notated $X_{(p)}$ is:

$$X_{(p)} = \{x \; s.t. \int_{-\infty}^{x} f(u)du = p\} \tag{15}$$

Here, for continuous r.v's, the *mean*, μ, and *variance*, σ^2 are parameters which we calculate from the probability density function. For continuous r.v., we weight each value by its probability $f(x)$ and use the integral for calculation.

Population Mean. Given a random variable X, defined on the real line $\mathbb{R} =]-\infty, \infty[$, the expectation is

$$\mu = \int_{\mathbb{R}} xf(x)dx = \mathbf{E}(X) \tag{16}$$

is the *mean* or *expectation* or *expected value* of X. In general:

$$\mathbf{E}(h(x)) = \int_{\mathbb{R}} h(x)f(x)dx \tag{17}$$

for any function, h.

Additionally, this equation — known as the *layered representation* — holds for non-negative random variables.

$$\mathbf{E}(X) = \int_{R^+} (1 - F(x))dx \tag{18}$$

Population Variance.

$$\sigma^2 = \mathbf{E}[(X - \mu)^2] = \int (x - \mu)^2 f(x)dx = Var(X) \tag{19}$$

Again, this can be reduced to:

$$\sigma^2 = Var(X) = E(X^2) - \mu^2 \tag{20}$$

Example: Say $X \sim f(x) = \frac{x}{2}$. Then

$$Var(X) = \int_0^2 (x - 8/6)^2 \cdot \frac{x}{2} dx =$$
$$= 8/9$$

General Joint Distributions

Two given random variables X and Y have a general, *joint distribution* that is an extension of the single variable definition. In the discrete case

$$\mathbf{P}((X, Y) = (x, y)) = p(x, y) \tag{21}$$

the *joint probability mass function*. In the continuous case

$$\mathbb{P}((X, Y) = (x \pm \epsilon, y \pm \epsilon)) = f(x, y) \tag{22}$$

We generate the *marginal* distributions for X and Y alone just as we did for contingency tables by summing over all values of the *other* variable.

$$p_x(x) = \sum_y p(x, y); p_y(y) = \sum_x p(x, y) \tag{23}$$

$$f_x(x) = \int_{\mathbb{R}} f(x, v) dv; f_y(y) = \int_{\mathbb{R}} f(u, y) dx \tag{24}$$

Two random variables are *independent* if

$$p_{X,Y}(x, y) = p_X(x) p_Y(y) \tag{25}$$

or

$$f_{X,Y}(x, y) = f_X(x) f_Y(y) \tag{26}$$

This is directly analogous to the general probability rules above. The conditional probability mass and density functions, are then just:

$$p_{X|Y}(X|Y) = \frac{p_{X,Y}(x, y)}{p_Y(y)} \tag{27}$$

$$f_{X|Y}(X|Y) = \frac{f_{X,Y}(x, y)}{f_Y(y)} \tag{28}$$

Mean of a Linear Transform. If X has mean μ_x, and a, b are constants (known), then:

$$\mathbf{E}(aX+b) = \int (ax+b)f(x)du = \int axf(x)du + \int bf(x)du = a\mathbf{E}(X) + b \tag{29}$$

If Y has mean μ_y, and c, d are constants (known) as well, then a natural extension to the single variable general case is:

$$E(aX+b+cY+d) = a\mu_x + bc\mu_y + d \tag{30}$$

Same as the discrete case. Notice the variance for the general linear transform: aX + b.

$$Var(aX+b) = a^2(\mathbf{E}(X^2) - {\mu_x}^2) = a^2 Var(X) \tag{31}$$

Since $\mathbf{E}((ax+b)^2) - [\mathbf{E}[(ax+b)]]^2 = a^2\mathbf{E}(X^2) + 2ab\mu_x + b^2 - a^2{\mu_x}^2 - 2ab\mu_x - b^2$. Again same as the discrete case.

Expectation and Covariance. The main idea is that expectation is a linear operator and that the expectations of a function is the expectation taken over the values of the function. In a natural extension to two dimensions:

$$E(g(X,Y)) = \begin{cases} \sum_y \sum_x g(x,y)p(x,y), & x,y\ discrete \\ \int\int g(x,y)f(x,y)dxdy, & x,y\ cont. \end{cases}$$

Example. We get the *moments* we use in calculation of mean and variance, etc. by choosing the function we take an expectation of.

$$g_1(x) = x \longrightarrow E(g_1(X)) = \mu_X$$

$$g_2(x,y) = xy \longrightarrow E(g_2(X,Y)) = E(XY)$$

$$g_3(y) = (y-\mu_y)^2 \longrightarrow E(g_3(Y)) = E((Y-\mu_y)^2) = Var(Y)$$

Covariance. Let $g(X,Y) = [X-\mu_X][Y-\mu_y]$. Then:

$$E(g(X,Y)) = E([X-\mu_X][Y-\mu_y])$$

$$= E(XY - \mu_Y X - \mu_X Y + \mu_X\mu_Y)$$

$$= E(XY) - \mu_Y E(x) - \mu_X E(Y) + \mu_X\mu_Y$$

$$E(XY) - \mu_X\mu_Y$$

This expectation has a special name, the *covariance* of X, Y. So the covariance of X, Y is

$$Cov(X,Y) = E([X-\mu_X][Y-\mu_Y]) = E(XY) - \mu_X\mu_Y \tag{32}$$

Properties of Covariance

• Covariance can be negative although $Var(X) \geq 0$

$$Cov(X,Y) \in \mathbb{R}$$

• Independence implies zero Covariance: If $X \perp Y$ then

$$E(XY) - \mu_X \mu_y = \mu_X \mu_Y - \mu_X \mu_Y = 0$$

• Zero Covariance does not imply independence: The fact here is $Cov(X,Y) \not\Rightarrow X \perp Y$.
For an example, take X, Y with this distribution:

$$\mathbb{P}(X=0) = \mathbb{P}(X=1) = \mathbb{P}(X=-1) = \frac{1}{3}$$

$$Y = \begin{cases} 0, & X \neq 0 \\ 1, & X = 0 \end{cases}$$

Thus $E(X) = 0$ and $E(XY) = 0$ but Y is obviously a function of X.

In general, for many $Y = g(X)$, where g is symmetric (about zero, for instance), $Cov(X,Y) = 0$ but X is — of course — not independent of $Y = g(X)$.

• Covariance is symmetric

$$Cov(X,Y) = Cov(Y,X)$$

$$Cov(X,X) = Var(X)$$

$$Cov(X,X) = E([X-\mu_X][X-\mu_X]) = Var(X)$$

• Covariance of linear transforms

$$Cov(ax+b, Y) = aCov(X,Y)$$

• Covariance of sum is sum of covariance: This is a generalization of the above:

$$Cov(\sum_i X_i, \sum_j Y_j) = \sum_i \sum_j Cov(X_i, Y_j)$$

• Variance of a sum of nonindependent random variables

Recall, if $X_1, ..., X_n \sim \mu_X, \sigma_X^2$ are independent then $Var(\sum_i X_i) = \sum_i Var(X_i)$. But, say $X_1, ..., X_n$ are <u>not</u> independent, i.e. $Cov(X_i, X_j) \neq 0$, then

$$Var(\sum_i X_i) = Cov(\sum_i X_i, \sum_j X_i)$$

$$= \sum_{i,j} Cov(X_i, X_j) = \sum_i Cov(X_i, X_i) + \sum_{i \neq j} Cov(X_i, X_j)$$

$$\Rightarrow Var(\sum_i X_i) = \sum_i Var(X_i) + 2\sum_{i<j} Cov(X_i, X_j)$$

and for a linear transformation

$$Var(\sum_i a_i X_i + b_i) = \sum_i a_i^2 Var(X_i) + 2\sum_{i<j} a_i a_j Cov(X_i, X_j)$$

The number ρ, which we introduced as a parameter to the multivariate normal distribution, is called the *correlation coefficient*

$$\rho = \frac{Cov(X,Y)}{\sqrt{\sigma_X^2 \sigma_Y^2}} = \frac{E([X-\mu_x][Y-\mu_Y])}{\sqrt{E([X-\mu_X]^2)E([Y-\mu_Y]^2)}} \tag{33}$$

Fact (a version of the Cauchy-Schwarz inequality):

$$|E([X-\mu_x][Y-\mu_Y])| \leq \sqrt{E([X-\mu_X]^2)E([Y-\mu_Y]^2)}$$

so $\rho \in [-1,1]$

Remark:

$$E(XY) = \mu_X \mu_Y + \rho \sigma_X \sigma_Y$$

since $Cov(X,Y) = \rho \sigma_X \sigma_Y$

Variance, Again!

This is important for a general equation for variance of linear transforms:

$$Var(aX+bY) = a^2 Var(X) + b^2 Var(Y) + 2abCov(XY) \tag{34}$$

Important Continuous Random Variables

Normal Distribution

When we write

$$X \sim N(\mu, \sigma^2)$$

we say that 'X is ***Normally distributed*** *with mean parameter mu and variance parameter sigma squared*'. The **parameters** for the Normal distribution are the mean μ and the variance σ^2.

The pdf for any Normal random variable, X, with parameters μ, σ^2 is

$$f(x) = (2\pi\sigma^2)^{-1/2} \exp\{-\frac{(x-\mu)^2}{2\sigma^2}\} \tag{35}$$

is the density function for the normal random variable.

This distribution, the *Normal Distribution*, is used as a model for continuous data that we believe to be unimodal and roughly symmetric.

You see here that it arises naturally as the distribution of sample averages — via the Central Limit Theorem. It arises naturally and often, in many circumstances.

So if someone writes $\mathcal{N}(5,25)$, you know to refer to a Normal distribution of mean 5 and variance 25 (standard deviation 5).

Linear Property of Normal. It turns out that if $X \sim (\mu, \sigma^2)$ then $Y = aX + b$, a, b both constants, then $Y \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$.

Standard Normal. The Standard Normal random variable is a special case - we set $\mu = 0$ and $\sigma^2 = 1$. We usually write

$$Z \sim N(0, 1)$$

reserving Z as a the special letter for the standard normal random variable.

Gamma Distribution

Continuous Time Factorial. We know that $n! = n(n-1)(n-2)(n-3)\cdots 1$ - this is the number of distinct arrangements of n distinct objects - taken all at a time. This notation and definition works well when n is an integer real number. Here are a few facts about the Gamma Function, some naturally from its definition as the factorial for integer real arguments.

- $\Gamma(z) = (z-1)\Gamma(z-1)$
- When n is an integer: $\Gamma(n) = (n-1)!$
- $\Gamma(1) = 1, \Gamma(1/2) = \sqrt{\pi}$
- $\Gamma(z)\Gamma(1-z) = \frac{\pi}{sin\pi z}$

General Form of the Gamma Distribution. Let's change a variable: $x = \beta t$ then, since $dx = \beta dt \rightarrow dt = \frac{dx}{\beta}$:

$$\Gamma(z) = \int_0^\infty (\frac{x}{\beta})^{z-1} e^{-x/\beta} \frac{1}{\beta} dx$$

Rearranging things, and calling $z = \alpha$

$$f(x; \alpha, \beta) = \frac{x^{\alpha-1} e^{-x/\beta}}{\Gamma(\alpha)\beta^\alpha} dx \tag{36}$$

We say X has a Gamma distribution with parameters α and β. $X \sim \Gamma(\alpha, \beta)$.

The Gamma distribution arises as the distribution of waiting times until (between) α number Poisson distributed events, each with parameter λ. We set $\beta = \lambda^{-1}$.

The Gamma distribution is useful because of its flexibility. Many non-negative processes can be modelled.

Exponential Distribution. A special case of the Gamma distribution is when $\alpha = k = 1$ and $\beta = \frac{1}{\lambda}$ — we call this the Exponential distribution, $X \sim Exp(\lambda)$.

$$f(x) = \lambda e^{-\lambda x} \tag{37}$$

The above is the pdf for the exponential distribution. Below

$$F(x) = 1 - e^{-\lambda x} \tag{38}$$

is the cumulative distribution function. So then, for $X \sim Gamma(\alpha = 1, \beta = \frac{1}{\lambda})$ we say $X \sim Exp(\lambda)$ so $\mathbf{E}(X) = \alpha\beta = \frac{1}{\lambda}$ and $Var(X) = \alpha\beta^2 = \frac{1}{\lambda^2}$

One reason the Exponential distribution is so important is its "memory-less" property. That is, the distribution of successive intervals of waiting time are independent of past waiting time. To see this, take $X \sim Exp(\lambda)$ and look at the probability one has to wait s units of time to see a Poisson arrival, given that t_o units of time have expired.

Chi-Squared Disribution. If we take $X \sim \Gamma(\alpha = \frac{r}{2}, \beta = 2)$ then

$$F(x) = \int_0^x \frac{1}{\Gamma(r/2)2^{r/2}} t^{r/2-1} e^{-t/2} dt \tag{39}$$

is the cumulative distribution for a strictly non-negative random variable that we call "Chi-squared". In notation we say $X \sim \chi^2(r)$ or "X has a chi-square distribution with r degrees of freedom".

This distribution is widely used and is tabled in the back of most introductory statistics textbooks. There is a specific use of the Chi-squared distribution, which illustrates its wide application

Squared Deviations are Chi-Squared distributed. Take $X \sim N(\mu, \sigma^2)$. We know we can "standardize" it by setting $Z = \frac{X-\mu}{\sigma}$. $Z \sim N(0, 1)$, which you should be able to verify!

Z is an absolute measure of the deviation of X, scaled by σ. We know that, on average $\mathbf{E}(Z) = 0$, which may obscure the magnitude of many deviations (some will be positive, others negative – they will cancel). When it is important, we construct $Z^2 = \frac{(X-\mu)^2}{\sigma^2}$ as a measure of the squared deviation. Then $V = Z^2$ has a Chi-Squared distribution.

t- Distribution

If $X \sim \mu, \sigma^2$ is a random variable [model], and we collect some data $\overline{x} = \sum_i^n x_i$, then the distribution of the sample mean[2], when the variance is known is $\overline{x} \sim N(\mu, \sigma^2/n)$.

We often construct the *test statistic*

$$Z = \frac{\overline{X} - \mu_0}{\sqrt{\sigma^2/n}}$$

which has the standard normal distribution $N(0, 1)$. The Z statistic is the deviation of the data from the null hypothesis, over its standard deviation. In words:

$$Z \equiv \frac{obs - exp}{S.D(obs)}$$

is the statistic we want to use if we want to make a probabilistic statement about the true mean μ, using observed data $X_1, ..., X_n$, in many situations we cannot assume that we know the variance of the sample mean. As well, we often have not enough samples to apply the central limit theorem to the sampling distribution. In these situation we construct the *t-statistic*:

$$t = \frac{\overline{X} - \mu}{s/\sqrt{n}} \tag{40}$$

[2]See sampling, below

The t-distribution, $T \sim t(df)$ is an approximation to the normal distribution. Notice I have written df as the *parameter* of the distribution.[3] The T distribution is centered at zero, just like the Z.[4]. We let $df \equiv degrees\ of\ freedom$.

When we talk about sample data, we loosely define "degrees of freedom" as the number of independent observations — *the number of observations we have left after we* ***subtract the number of parameters we have to estimate***.

$$df \equiv n - k$$

where we let $n =$ the number of observations and $k =$ the number of parameters to be estimated.

More generally,

$$T = \frac{Z}{\sqrt{V/r}}$$

with Z and V independent, the density function for the t distribution is

$$f_{Z,V} = \frac{1}{\sqrt{2\pi}e^{-z^2/2}} \cdot \frac{1}{\Gamma(r/2)2^{r/2}} v^{r/2-1}e^{-v/2}$$

F-distribution for ratio of variance. If $X_1, ...X_m$ is distributed $N(\mu_1, \sigma_1^2)$ and $Y_1, ...Y_n$ is distributed $N(\mu_2, \sigma_2^2)$ then the ratio

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \tag{41}$$

has what we call an F distribution with *numerator degrees of freedom* $m - 1$ and *denominator degrees of freedom* $n - 1$.

F is the ratio of two independent chi-squared variables, call them $U \sim \chi^2(m-1)$ and $V \sim \chi^2(n-1)$. If $U = \frac{(m-1)S^2}{\sigma_1^2}$ then $U \sim \chi^2(m-1)$.

Statistical Estimation

The sampling distribution

Let's make this assumption. We have sampled some data, $x_1, ..., x_n$ from a population that (we believe) is distributed $X \sim \mu, \sigma^2$.

Remember that $\overline{X} = \sum \frac{X_i}{n} = \frac{1}{n}\sum X_i$. So, using what we know about expectations:

$$E(\overline{X}) = E(\sum \frac{X_i}{n})$$

$$= \frac{1}{n}E(\sum X_i) =$$

$$= \frac{1}{n}[E(X_1) + \cdots + E(X_n)]$$

[3]What, if any, are the parameters for the Z $N(0,1)$ distribution? The parameters are $\mu = 0$ and $\sigma^2 = 1$.

[4]It turns out that $\mathbf{E}(T) = 0$ and $Var(T) = \frac{r}{r-2}$

$$= \frac{1}{n}[n\mu]$$

$$E(\overline{X}) = \mu \tag{42}$$

In words: *the expectation of the sample mean is the population mean*.

As well, using what we know about the variance:

$$Var(\overline{X}) = Var(\sum \frac{X_i}{n})$$

$$= \frac{1}{n^2} Var(\sum X_i) =$$

$$= \frac{1}{n^2}\sigma^2$$

$$Var(\overline{X}) = \frac{\sigma^2}{n} \tag{43}$$

In words: *the variance of the sample mean is the population variance divided by the size of the sample*.

These statements, equations (42) and (43) are true for any random variable ($X \sim \mu, \sigma^2$). These are the *mean and variance of the sample mean*. Although equations (42) and (43) are general, the distribution of $\bar{X}$ depends on the distribution of the sample data $X_1, \ldots, X_n$. This is called the *sampling distribution*.

Conditional Expectation and Prediction. The notion of conditional expectation, expectation of a random variable (or process), given additional information — i.e. another random variable. In this class, and in almost every other statistics class — the conditional expectation is exploited in the *regression* setup. A regression is usually a conditional expectation. In a sense, we describe or *predict* random variables by their expectations; a regression is a conditional prediction.

Estimators and Estimation

In general, *point estimation* is the procedure of selecting a "best" value for a parameter, and thus specifying a probability model. The setup, using mathematical notation, is like this. Let a random variable X have a pdf $f(x;\theta)$. Some Examples we have seen: $X \sim Poi(\lambda)$ so $f(x;\theta = \lambda) = \frac{\lambda^x e^{-\lambda}}{x!}$; $X \sim N(\mu, \sigma)$ so $f(x;\theta_{(\mu,\sigma^2)}) = (2\pi\sigma^2)^{-1/2} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$; etc. The procedure of *point estimation* is using a function of the data $\mathbf{X} = (x_1, \ldots, x_n)$, as a statistic which will be a good estimator of the parameter(s) θ.

Bias

"Theta hat" or $\widehat{\theta}$ is our estimator for θ and we call the *bias* the difference between the expected value of the estimator and the parameter being estimated:

$$Bias = \mathbf{E}(\widehat{\theta}) - \theta \tag{44}$$

We call an estimator *unbiased* if the bias is zero. For example, $S^2 = (n-1)^{-1}\sum_i^n (x_i - \overline{x})^2$ is an unbiased estimator for σ^2 since $\mathbf{E}(S^2) = \sigma^2$.

Maximum Likelihood Estimation

Say we have a random variable, random process, whatever: $X \sim f(x;\theta)$ and a sample $x_1, ..., x_n$, simple and random, each $x_i \sim f(x_i;\theta)$. In words, we have some random process with a pdf (distribution) and a sample where each of the observations has the same pdf – θ is the value of the parameter that specifies the exact distribution from the family $f(x)$.

The probability of observing a particular sample $x_1, ..., x_n$ is the joint probability of $x_1, ..., x_n$. We call this the *likelihood*:

$$lik(\theta; \mathbf{x}_n) = f(\mathbf{x}_n; \theta) \tag{45}$$

Notice that here we are considering the *likelihood* as a function of parameter θ: the likelihood is the probability that we would observe a sample at a value of θ.

When we have a simple random sample, we assume each observation in the sample is distributed independently and equivalently: We can rewrite the likelihood:[5]

$$lik(\theta; \mathbf{x}_n) = \prod_i^n f(x_i; \theta) \tag{46}$$

Now, we just need to pick a value of θ such that (46) is maximized. So we turn the estimation problem into a calculus problem: we have to maximize a function.

Often we take the logarithm of the likelihood function[6].

$$log(lik(\theta; \mathbf{x}_n)) = log(\prod_i^n f(x_i; \theta)) = \sum_i^n log(f(x_i; \theta)) \tag{47}$$

There are at least two good reasons for doing this. First, taking the logarithm allows us to consider a sum instead of a product. We like sums[7]. Second, taking the logarithm introduces concavity. This allows us to find the value of the parameter $\theta = \theta*$ which maximizes the log likelihood. This value $\theta*$ we will call the *maximum likelihood estimate* for θ and use it as our estimate $\widehat{\theta}$.

We'll talk about the *method of least squares* when we talk about linear regression, soon.

Statistical Inference

The general setup

Take a random variable $X \sim \mu, \sigma^2$ and with n very very large.
Then $\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$.

[5]Product notation: Remember the summation notation? The product notation is the same, with a $*$ separating each term instead of a $+$

[6]I always write $log()$. I always mean $ln() = log_e()$

[7]The CLT refers to sums, not products

We make

$$Z_o = \frac{\overline{X} - \mu_{\overline{X}}}{\sigma_{\overline{X}}}$$

our ***observed z-statistic*** and we ask:

What are the interval limits for μ at a given probability? This is the same thing as asking, after rescaling to Z, what are the interval limits for a Z at a given probability, say .95? Using our notation this looks like:

$$\mathbb{P}(-1.96 \leq Z_o \leq 1.96) = .95$$

Using algebra and some maneuvers,

$$\mathbb{P}(-1.96 \leq \frac{\overline{X} - \mu_{\overline{X}}}{\sigma_{\overline{X}}} \leq 1.96) = .95$$

$$\mathbb{P}(\overline{X} - 1.96 * \sigma_{\overline{X}} \leq \mu_{\overline{X}} \leq \overline{X} + 1.96 * \sigma_{\overline{X}}) = .95$$

In words, the probability that the true mean is within 1.96 times the standard deviation of the sample mean is 95 percent. Thus, the ***95 percent confidence interval*** for the true mean is 1.96$*$the standard deviation.

Notice how we set the confidence level we wanted (95 percent). Notice how this set the upper and lower limits for Z to be 1.96 and -1.96. In general, and $(1-\alpha)$ ***level confidence interval for*** μ ***the population mean*** is:

$$\overline{X} \pm Z_{\alpha/2} * \sigma_{\overline{X}}$$

In "ideas" the confidence interval always looks something like:

$$point\ estimate \pm Probability\ Distribution\ ptile * \ SD\ of\ point\ estimate$$

With a point estimate and a distribution for it we can usually generate a confidence interval for the parameter we seek to estimate.

Hypothesis Testing

The hypothesis testing procedure is as follows. We suppose a fixed value for the parameter of interest, say μ_0 and we set our null hypothesis, H_o, to be the statement that the true value of the parameter is this fixed value.

$$H_0 : \mu = \mu_0$$

We set two-sided, or one-sided alternatives to this hypothesis, H_a.

$$H_a : \mu \neq \mu_0$$

or

$$H_a : \mu > \mu_0$$

or

$$H_a : \mu < \mu_0$$

We use our test statistic, that is the rescaled distribution of the sample estimate, to determine which hypothesis to accept. Here we are testing the population mean, μ. The estimate of the population mean is the sample mean, $\overline{x}$. We know the distribution of the sample mean is $\overline{x} \sim N(\mu, \sigma^2/n)$ so we use a z-statistic

$$Z = \frac{\overline{x} - \mu_0}{\sqrt{\sigma^2/n}}.$$

The z-statistic is the distance, in standard normal units, of what we observed, $\overline{x}$, from what we expect to observe under the null hypothesis, μ. We divide the deviation, $\overline{x} - \mu$ by standard deviation of our estimate - $\sqrt{\sigma^2/n}$. Look at the footnote.[8]

We complete the hypothesis test by computing the probability we would have seen such an extreme result in the data under the assumptions of the null hypothesis. This is called the p-value.

$$\mathbb{P}(Observed\, Test\ Statistic,\ given\ H_0\ is\ true)$$

If the p-value is bigger than the ***significance level*** for the test α, we cannot reject the null hypothesis. (This is actually the way we say it!). If the p-value is smaller than α, we reject the null hypothesis in favor of the alternative.

Wordy, but efficiently minimal statements that rely on the distribution of the sample estimate which, by the power of the central limit theorem, tends to be Normal.

Two-sided vs One-sided Hypothesis Tests

Notice that we can see the ***test statistic*** as the magnitude of the deviation of our sample from the null hypothesis, H_0. This deviation, measured by the magnitude of the test statistic, **need not be towards the alternative hypothesis** in question.

Two-sided test. Say we have a hypothesis test

$$H_0 : \mu = \mu_0\ vs.\ H_a : \mu \neq \mu_0.$$

A sample of data is taken $x_1, ..., x_n$ of which we take the sample mean $\overline{x}$. We construct the appropriate test statistic, **from the distribution of the estimate**. Here, of course, the sample mean is distributed $N(\mu, \frac{\sigma^2}{n})$. Our test statistic

$$Z = \frac{\overline{x} - \mu}{\frac{\sigma^2}{n}}$$

then, is the rescaling of the sample mean, and thus has the distribution $N(0, 1)$.

In this setup, where the alternative hypothesis is ***two-sided*** any large deviation from the null hypothesis is evidence for the alternative. Thus, a large magnitude test statistic z_o, positive or

[8] Test statistics have a general form of $test\ statistic = \frac{observed\ value - expected\ value}{std\ dev.\ of\ estimate}$

negative, is evidence for rejection of the null hypothesis. In this, the two sided hypothesis test, the p-value is:

$$\mathbb{P}(Z \geq |z_o|)$$

since we look at both sides of the distribution for evidence against the null hypothesis.

One-sided test. In the ***one-sided test*** the alternative hypothesis considers only departures from the null in one direction. These hypothesis tests:

$$H_0 : \mu = \mu_0 vs. H_a : \mu > \mu_0$$

or

$$H_0 : \mu = \mu_0 vs. H_a : \mu < \mu_0$$

are examples of 1 tailed tests. For each of the tests, only a large deviation – i.e. a large magnitude test statistic – *in the direction of the alternative hypothesis* is evidence against the null. ***A large test statistic not in the direction of the alternative is not*** evidence against the null hypothesis here.

P-value

We've already defined the ***p-value*** as a measure of how plausible the null hypothesis H_0. Say we conduct a 1-tailed test, using z_o as our test statistic. Then, our p-value is

$$\mathbb{P}(Z > z_o)$$

or

$$\mathbb{P}(Z < z_o)$$

depending if our alternative hypothesis looks for deviations above or below, respectively. Notice that in the first case z_o is often positive; we look for large positive deviations from the null. In the second z_o is often negative; we look for large negative deviations from the null

For a two-sided test, again using z_o as our test statistic, the p-value is

$$\mathbb{P}(Z > |z_o|)$$

Here, any large deviation for the null is evidence for the alternative hypothesis. In general, when the p-value is small we reject the null hypothesis.

We can also say that the p-value *is the smallest significance level at which we reject the null hypothesis*. For a 1-tailed test, the rejection region is far above or below the null hypothesis – determined by the significance level of the test. For a two-sided test, the rejection region is both above and below the null hypothesis; deviations in either direction are cause for rejection of the null hypothesis.

Testing errors

We have been calling α the ***significance level***; $1-\alpha$ is the familiar ***confidence level***.

Remark:

$$\alpha = \mathbb{P}_{H_0}(A\, test\ statistic\, is\, in\, rejection\, region) \tag{48}$$

That's why when the $p-value < \alpha$ we reject the null hypothesis. To restate, when the probability any test statistic is in the rejection region is greater than the probability of our observed statistic — we reject the null hypothesis.

On the other hand

$$\beta = \mathbb{P}_{H_a}(A\, test\ statistic\, is\, NOT\, in\, rejection\, region) \tag{49}$$

Notice that both α and β are probabilities of making a hypothesis testing error – choosing the wrong hypothesis when the other is true.

- α is commonly called ***type 1 error***: a rejection of the null when it is in fact true.
- β is commonly called ***type 2 error***: a selection of the null when it is in fact false.

In these hypothesis testing setups you have but two choices and there are but four possible outcomes. You make the right choice (2), the correct decision (2), and what actually happens $2*2=4$. Below, is the same sort of table you would see in your text.

	The Truth	
Your Decision	$\mathbf{H_0}$	$\mathbf{H_a}$
$\mathbf{H_0}$	Confidence Level	β
$\mathbf{H_a}$	α	Power

Samples - Independent or Dependent?

In general always remember: ***(1) The sampling distribution***, which will yield the ***(2) The confidence interval***, which is immediately analogous to ***(3) The test statistic***. Everything is a variation on this theme, just a slightly different scenario.

Scenario 1: Two sample proportions

Say we wish to gain inference on the support for election reform in California and Georgia. Let $p_1 \equiv$ the proportion who support in Georgia and $p_2 \equiv$ the proportion who support in California. We estimate these, in the usual way, $\widehat{p}_1 = \frac{x_1}{n_1}, \widehat{p}_2 = \frac{x_2}{n_2}$: the sample proportions of voters who supported the reform over total voters, for each state.

We know from the sampling distribution of $\widehat{p}$: $E(\widehat{p}_1) = p_1, E(\widehat{p}_2) = p_2$ and $Var(\widehat{p}_1) = \frac{p_1 q_1}{n_1}, Var(\widehat{p}_2) = \frac{p_2 q_2}{n_2}$. The difference $p_1 - p_2$ is distributed:

$$p_1 - p_2 \sim N(p_1 - p_2, \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2})$$

This is ***the sampling distribution for the difference in proportions***. The appropriate rescaled statistic is:

$$Z = \frac{\widehat{p}_1 - \widehat{p}_2 - (p_1 - p_2)}{S.D.(\widehat{p}_1 - \widehat{p}_2)}$$

and it will have a standard normal distribution.

Thus, a confidence interval for the difference in two proportions is:

$$\widehat{p}_1 - \widehat{p}_2 \pm Z_{\alpha/2}\sqrt{\frac{\widehat{p}_1\widehat{q}_1}{n_1} + \frac{\widehat{p}_2\widehat{q}_2}{n_2}}$$

For the two tailed hypothesis test
$H_0 : p_1 = p_2$ vs. $p_1 \neq p_2$
we exploit the fact $p_1 = p_2$ implies $p_1 - p_2 = 0$ and write

$$\widehat{p}_{pooled} = \widehat{p}_p = \frac{x_1 + x_2}{n_1 + n_2}$$

to ***pool*** the estimate of the population proportion, since, under the null, here, $p_1 = p_2$.

Then our test statistic is

$$z_o = \frac{\widehat{p}_1 - \widehat{p}_2}{\sqrt{\widehat{p}_p\widehat{q}_p(\frac{1}{n_1} + \frac{1}{n_2})}}.$$

Scenario 2: Two samples, in general

In general if we have data coming from two samples $X_1 \sim \mu, {s_1}^2$ and $X_2 \sim \mu, {s_2}^2$ and we cannot assume knowledge of the variances we get a sampling distribution for the difference in the population mean $\mu_1 - \mu_2$ as

$$\overline{x}_1 - \overline{x}_2 \sim \mu_1 - \mu_2, \frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}$$

which we approximate with a t-distribution with $n_1 + n_2 - 2$ degrees of freedom.[9]

Thus the **confidence interval is**

$$\overline{x}_1 - \overline{x}_2 \pm t_{\alpha/2, n_1+n_2-2} * \sqrt{\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}}.$$

The two sided **hypotheses test for differences in the population mean** $H_0 : \mu_1 - \mu_2 = \Delta_0$ vs. $H_a : \mu_1 - \mu_2 \neq \Delta_0$ would use this test statistic:

$$t_0 = \frac{\overline{x}_1 - \overline{x}_2 - \Delta_0}{\sqrt{\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}}}.$$

[9]The exact calculation for degrees of freedom here is more involved. Using $n_1 + n_2 - 2$ is good.

Of course one sided tests are the usual variations on this.

If **you are willing to assume that** $s_1 = s_2$ then you can ***pool the variance estimates*** with

$$S_p{}^2 = \frac{(n_1 - 1){s_1}^2 + (n_2 - 1){s_2}^2}{n_1 + n_2 - 2}$$

and use this test statistic:

$$t_0 = \frac{\overline{x}_1 - \overline{x}_2 - \Delta_0}{\sqrt{S_p{}^2(\frac{1}{n_1} + \frac{1}{n_2})}}$$