Theory of Statistical Inference Notes

Shubhi Sharma

2/22/2020

Contents

Basic Probability Theory **3**

Random variables 3

Probability measures 3

Univariate Probability Distributions **4**

Properties of CDFs 4

Discrete CDFs 4

Continuous CDFs 5

Change of variables 5

Multivariate Probability Distributions **9**

Marginal distributions 9

Multivariate margins 10

Conditional distributions 10

Independence 11

Multivariate change of variables 11

Moments **13**

Expectation of a function 13

Properties of $\mathbb{E}[.]$ 13

Variance 14

Jenson's inequality 15

Covariance 15

Cauchy-Schwarz inequality 16

Conditional expectation 16

Inference with Sample Mean **17**

Sample mean estimator 17

Linear Shrinkage Estimator 17

Precision of $\hat{\theta}$ 17

Variance of sample mean 18

Confidence intervals 18

Markov inequality 19

Confidence interval via Markov's inequality 20

Chebyshev's inequality 20

Confidence interval via Chebyshev's inequality, ... 21

Normal confidence interval 21

Convergence in Probability **22**
Law of Large Numbers 22
Consistency 22
Central Limit Theorem 23
Equality in Distribution 23
Moment Generating Function 23
Convergence in Distribution 25

Consistent and Asymptotically Normal (CAN) Estimators **27**
Asymptotic Confidence Intervals 27
Delta method 28
Convergence in probability is convergence in distribution 28
Multivariate CLT and Delta Method 29

Parameteric Inference and Likelihood Methods **30**
Likelihood and MLEs 30
Consistency of MLE 31
M Estimation 32
Identifiability 33

Information and Asymptotic Normality **33**
Fisher Information 34
Variance of score information 34
Asymptotic normality 34

Hypothesis testing **34**
Statistical hypothesis testing 34
Level α tests 35
Hypothesis testing and CI 36
Randomization and Permutation tests 37
P-values 37
Multiple camparisons 38
Multigroup effects 38
Linear regression 39

Multiple Testing **39**
Global Null and Error 39
Global control vida Bonferroni's correction 40
Bonferroni Method 40
Limitations of Bonferroni 41
Fisher's combined probability test 41
False Discovery Rate 42

Basic Probability Theory

Random variables

Suppose you are going to perform a survey, or do an experiment, do a quantitative study of a population or process.

Let Ω be the set of possible outcomes of the study and let $w \in \Omega$ be the actual outcome. Since the study hasn't been done yet, w is unknown and so it is said to be "random".

Let Y be some numerical summary of your study outcome. Then Y is a function of w: $Y = Y(w)$. Since w is random, so is Y. It varies as a function of the random outcome w.

Probability measures

Let P be a probability measure over Ω, meaning,

$$0 = P(\phi) \leq P(A) \leq P(\Omega) = 1$$

and,

$$P(\cup_i^\infty A_i) = \sum_i^\infty P(A_i)$$

if $A_i, A_{i'}$ is disjoint.

Consider, $\mathbb{Y} \in \mathbb{R}$,

$$Pr(Y \leq y) = P(w : Y(w) \leq y) = P(Y_{(-\infty,y]})$$

This is why sometimes, "random variables are functions". An example of this, Consider a populations with a collection of people $\{1, 2, \cdots, N\}$. The sampling procedure proceeds by uniformly selecting a person at random. Unknown outcome is the unknown index w of the selected person $Pr(w = i) = P(\{i\}) = 1/N, \quad i = 1 \cdots N$.

Let Y represent the BMI of the ith person in the population. Then, $Y(w)$ is a random variable where w is the index of the randomly selected individual.

What is the induced distribution?

$$\begin{aligned} Pr(Y(w) \leq y) &= Pr(w \in \{i : Y(i) \leq y\}) \\ &= P(\cup_{i:Y(i)\leq y}\{i\}) \\ &= \sum_{i:Y(i)\leq y} P(\{i\}) \\ &= \sum_{i:Y(i)\leq y} \frac{1}{N} \\ &= \frac{\#(Y_i \leq y)}{N} \end{aligned}$$

Univariate Probability Distributions

Let,

$$\begin{aligned} Y &\subset R \\ (-\infty, y] \cap \mathbb{Y} &\in \mathbb{B} \text{ for all } y \in \mathbb{R} \\ P & \text{ be a probability measure on } (\mathbb{Y}, \mathbb{R}) \end{aligned}$$

Cumulative distribution function (CDF): The CDF of P or of a random variable with distribution P, is the function $F : \mathbb{R} \to [0, 1]$ given by,

$$F(y) = P((\infty, y] \cap \mathbb{Y}) = P(Y \leq y)$$

Properties of CDFs

$$\begin{aligned} &1. \quad \lim_{y \to -\infty} F(y) = 0 \\ &2. \quad \lim_{y \to \infty} F(y) = 1 \\ &3. \quad F(y) \leq F(y + \delta), \delta > 0 \\ &4. \quad \lim_{\delta \to 0} F(y + \delta) = F(y) \\ &5. \quad \lim_{\delta \to 0} F(y - \delta) \text{exists} \end{aligned}$$

This is to say the CDF is non-decreasing, right continuous and bounded. (CADLAC functions).

More importantly, the CDF determines the probability distribution of $Y \sim P$.

Theorem. If Y_1, Y_2 are two random variables and

$$F(y_1) \equiv Pr(Y \leq y_1) = Pr(Y \leq y_2) \equiv F(y_2)$$

then, $Pr(Y_1 \in A) = Pr(Y_2 \in A)$. Equal CDFs means equal probability measures.

Discrete CDFs

If $\mathbb{Y}$ is finite or countable, then F is piecewise constant, with jumps/discontinuities at $y \in \mathbb{Y}$ such that $P(\{y\}) > 0$.

PDFs from CDFs

$$\begin{aligned} F(y) = Pr(Y \leq y) &= Pr(Y < y) + Pr(Y = y) \\ Pr(Y = y) &= F(y) - Pr(Y < y) \\ Pr(Y = y) &= F(y) - \sup_{y' < y} F(y') \end{aligned}$$

This is the probability density function of Y when Y is a discrete variable.

Properties of PDF for discrete random variables

$$\begin{aligned}
&1. \quad 0 \leq p(y) \leq 1 \\
&2. \sum_{y in \mathbb{Y}} p(y) = 1 \\
&3. p(y) = F(y) - \sup_{y' < y} F(y')
\end{aligned}$$

Continuous CDFs

Note for continuous CDFs, the probability of any one number is 0!

Definition: For a continuous variable Y, the CDF is defined as,

$$Pr(Y \leq y) = F(y) = \int_{-\infty}^{y} p(y) dy$$

It is a function of the pdf.

Properties of PDF for continuous variables,

$$\begin{aligned}
&1. \quad 0 \leq p(y) \\
&2. \quad 1 = \int_{-\infty}^{\infty} p(y) dy \\
&3. \quad p(y) = \frac{d}{dy} F(y)
\end{aligned}$$

Keep in mind that the probability density function is **NOT** the probability that $Y = y$. It is possible that $p(y) > 1$, but not on any interval of length > 1.

Change of variables

Derivation

Let Y be a random variable with a known CDF and PDF. Define W as a function of Y, such that Y = g(W). To find the PDF of W, first write the CDF of W in terms of a known CDF (i.e. in terms of Y), and then differentiate.

$$
\begin{aligned}
F_w(w) &= Pr(W \leq w) \\
&= Pr(g^{-1}(Y) \leq w) \\
&= Pr(Y \leq g^{-1}(w)) \\
&= F_y(g^{-1}(w))
\end{aligned}
$$

Differentiate,

$$
\begin{aligned}
p_w(w) &= \frac{d}{dw} F_w(w) \\
&= \frac{d}{dw} F_y(g^{-1}(w)) \\
&= p_y(g^{-1}(w)) \frac{d}{dw} g^{-1}(w)
\end{aligned}
$$

When g is monotonic i.e. strictly increasing or strictly decreasing, the more general change of variable formula is,

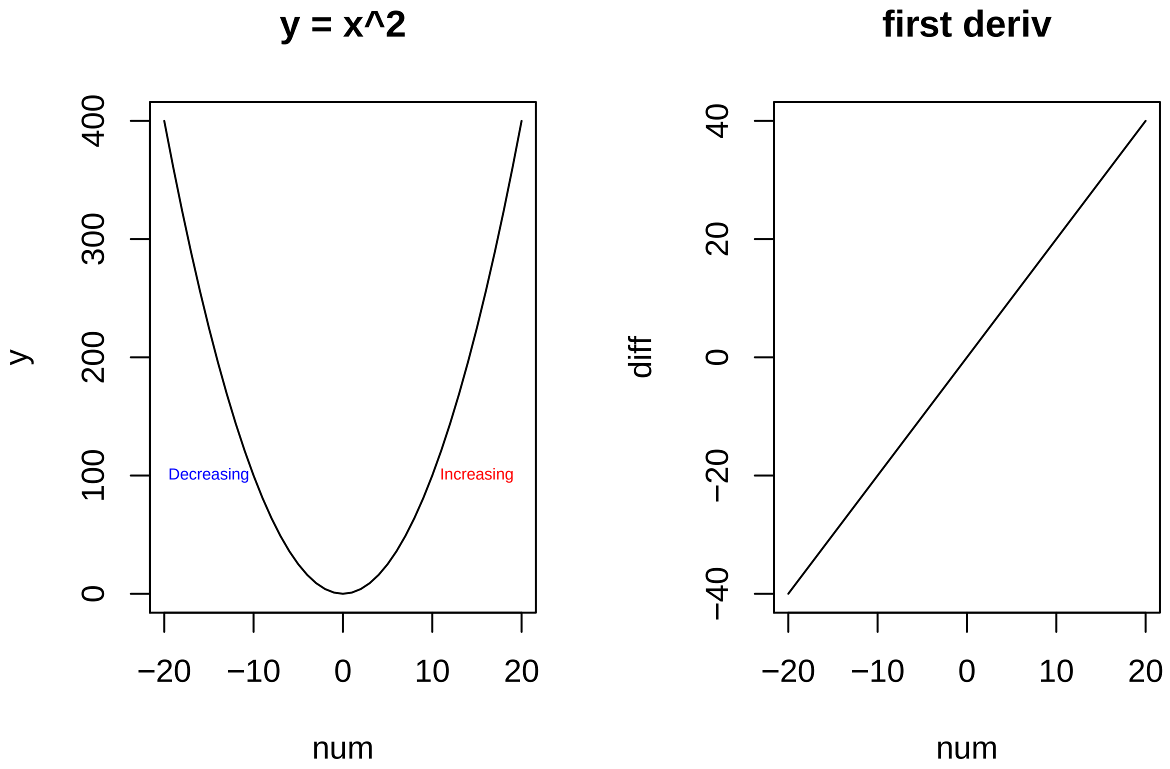
$$p_w(w) = p_y(g^{-1}(w)) \mid \frac{d}{dw} g^{-1}(w) \mid$$

What happens when g is not monotonic?

Example: $Y \sim N(0,1)$. What is the PDF of $X = Y^2$? Here, the function is not monotonic. Why? A function is monotonic if the sign of it's fist derivative does not change.

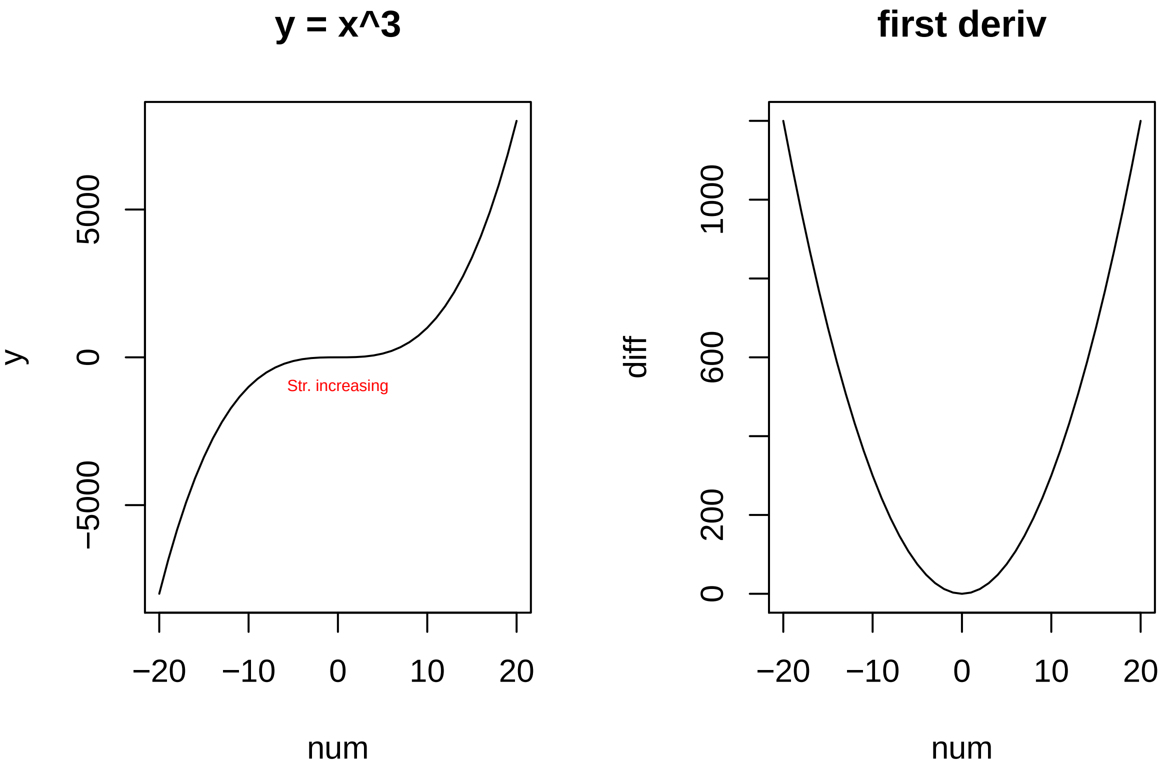
$$
\begin{aligned}
\frac{d}{dx} x^2 &= 2x^{2-1} = 2x \\
\frac{d}{dx} (-x)^2 &= -2x
\end{aligned}
$$

```
num <- seq(-20, 20, by = 1)
y <- num^2
diff <- 2*num

par(mfrow= c(1,2))
plot(num, y, type = "l", main = "y = x^2")
text(-15, 100, label = 'Decreasing', cex = 0.5, col = 'blue')
text(15, 100, label = 'Increasing', cex = 0.5, col = 'red')
plot(num, diff, type = "l", main = "first deriv")
```



An example of a monotonic function would be $g(x) = x^3$

```
num <- seq(-20, 20, by = 1)
y <- num^3
diff <- 3*num^2

par(mfrow= c(1,2))
plot(num, y, type = "l", main = "y = x^3")
text(0, -1000, label = 'Str. increasing', cex = 0.5, col = 'red')
plot(num, diff, type = "l", main = "first deriv")
```



Returning to the example, to account for the fact that $f(x) = x^2$ is not a monotnoic function, we have to take into account the probability when x is increasing (i.e. is +ve) and when x is decreasing (i.e. is -ve). In other words, we are splitting the funciton up for when x is str. increasing and when x is str. decreasing.

$$\begin{aligned}
F_x(x) &= Pr(X \leq x) = Pr(Y^2 \leq x) \\
&= Pr(-\sqrt{x} \leq Y \leq \sqrt{x}) \\
&= Pr(Y \leq \sqrt{x}) - Pr(Y \leq -\sqrt{x}) \\
&= F_y(\sqrt{x}) - F_y(-\sqrt{x}) \\
p_x(x) &= \frac{d}{dx}F_y(\sqrt{x}) - \frac{d}{dx}F_y(-\sqrt{x}) \\
&= p_y(\sqrt{x})\frac{x^{-1/2}}{2} - p_y(-\sqrt{x})\left(\frac{-x^{-1/2}}{2}\right) \\
&= p_y(\sqrt{x})x^{-1/2} \\
&= \frac{1}{\sqrt{2\pi}}x^{-1/2}e^{-x/2} \\
&\sim \chi^2 \text{ density}
\end{aligned}$$

Multivariate Probability Distributions

In general, we are more interested in more than one variable for example, we are interested in $Pr(\{Y_1, \cdots, Y_n\} \in B$. Formally, we say random variable on a shared proabbility space i.e. $Pr(\{w : Y_1(w) = y_1, \cdots, Y_n(w) = y_n\} \in B)$

Definition: In the discrete case, if all Y_i's are discrete, the joint pdfs as

$$\begin{aligned} P(y_1 \cdots y_n) &= Pr(Y_1 = y_1 \text{ and } Y_2 = y_2, \cdots, \text{ and } Y_n = y_n) \\ &= Pr(\{w : Y_1(w) = y_1\} \cap \cdots \cap \{Y_n(w) = y_n\}) \end{aligned}$$

Properties

$$\begin{aligned} &1. \quad 0 \leq p(y_1, \cdots, y_n) \leq 1 \\ &2. \quad \sum_{y_1} \cdots \sum_{y_n} p(y_1, \cdots, y_n) = 1 \end{aligned}$$

Definition: In the continuous case, the joint probability measure is defined as $Pr(Y_1 \in A_1$, and $Y_2 \in A_2 \cdots$ and $Y_n \in A_n)$. While the joint probability density $p(y_1, \cdots, y_n)$ is defined as a function such that,

$$Pr(Y_1 \in A_1, \cdots, Y_n \in A_n) = \int_{y_1} \cdots \int_{y_n} p(y_1 \cdots y_n) dy_1 \cdots dy_n$$

with the following properties

$$\begin{aligned} &1. \quad 0 \leq p(y_1 \cdots y_n) \\ &2. \quad \int_{y_1} \cdots \int_{y_n} p(y_1 \cdots y_n) dy_1 \cdots dy_n = 1 \end{aligned}$$

Note, that the probability density function is not the probability that $Y = y$. Therefore, it is possible that $p(y) > 1$, but not on any interval on length >1.

Marginal distributions

In the general discrete case,

$$Pr(Y_1 = y_1) = \sum_{y2} \cdots \sum_{y_n} p(y_1, \cdots y_n) \equiv p(y_1)$$

where y_1 is fixed.

In the continuous case,

$$
\begin{aligned}
Pr(Y_1 \in B) &= Pr(Y_1 \in B, Y_2 \in \mathbb{Y}_2, \cdots, Y_n \in \mathbb{Y}_{\ltimes}) \\
&= \int_B \int_{\mathbb{Y}_{\nvDash}} \cdots \int_{\mathbb{Y}_{\ltimes}} p(y_1 \cdots y_n) dy_1 \cdots dy_n \\
&= \int_B p_1(y_1) dy_1
\end{aligned}
$$

where y_1 is fixed and all the other y's are integrated over.

Multivariate margins

$$
\begin{aligned}
Pr(Y_2 \in B_2, Y_3 \in B_3) &= Pr(Y_1 \in \mathbb{Y}_{\nVdash}, Y_2 \in B_2, Y_3 \in B_3, Y_4 \in \mathbb{Y}_{\ntrianglerighteq}) \\
&= \int_{B_2} \int_{B_3} \int_{\mathbb{Y}_{\nVdash}} \int_{\mathbb{Y}_{\ntrianglerighteq}} p(y_1, y_2, y_3, y_4) dy_1 dy_2 dy_3 dy_4 \\
&= \int_{B_2} \int_{B_3} \left(\int_{\mathbb{Y}_{\nVdash}} \int_{\mathbb{Y}_{\ntrianglerighteq}} p(y_1, y_2, y_3, y_4) dy_1 dy_4 \right) dy_2 dy_3 \\
&= \int_{B_2} \int_{B_3} p(y_2, y_3) dy_2 dy_3 \\
&= p_{23}(y_2, y_3)
\end{aligned}
$$

Conditional distributions

The conditional probability of B given A is defined to be,

$$
P(A \mid B) = \frac{P(A \cap B)}{P(B)}
$$

In the discrete case, this translates to,

$$
\begin{aligned}
\frac{Pr(X \in A \cap Y \in B)}{Pr(Y \in B)} &= \frac{\sum\limits_{X \in A} \sum\limits_{Y \in B} p(x, y)}{\sum\limits_{x \in X} \sum\limits_{Y \in B} p(x, y)} \\
&= \frac{\sum\limits_{X \in A} \sum\limits_{Y \in B} p(x, y)}{\sum\limits_{y \in B} p(y)}
\end{aligned}
$$

In the continuous case, suppose X and Y have a joint continuous distribution,

$$
\begin{aligned}
P(X \in A \mid Y \in B) &= \frac{Pr(X \in A \cap Y \in B)}{Pr(Y \in B)} \\
&= \frac{\int_A \int_B p_{xy}(x, y) dx dy}{\int_B p_y(y) dy}
\end{aligned}
$$

For each $Y \in \mathbb{R}$, there is a probability density given by,

$$Pr(X \in B \mid Y = y) = \int_B p(x \mid y)dx$$

This is the conditional density given that $Y = y$.

How does $p_{x|y}(x \mid y)$ correspond to $Pr(X \in A \mid Y \in B)$?

Let $B_\epsilon = (y, y + \epsilon)$,

$$\begin{aligned}
\lim_{\epsilon \to 0} Pr(X \in A \mid Y \in B_\epsilon) &= \lim_{\epsilon \to 0} \frac{Pr(X \in A \cap Y \in B_\epsilon)}{Pr(Y \in B_\epsilon)} \\
&= \lim_{\epsilon \to 0} \frac{\int_A \int_y^{y+\epsilon} p(x, y)dxdy}{\int_y^{y+\epsilon} p(y)dy}
\end{aligned}$$

Since this limit approaches 0, using L'Hospital's Rule,

$$\begin{aligned}
&= \lim_{\epsilon \to 0} \frac{\int_A p(x, y)dxdy}{p(y)} \\
&= \frac{\int_A p(x, y)dxdy}{p(y)} \\
&= \int_A p(x \mid y)dxdy
\end{aligned}$$

In summary, for a discrete distribution, the conditional pdf is the joint over the marginal. For continuous variables, the conditional pdf can be derived from the limit of the conditional probability.

Independence

Definition: Events $A_1, \cdots A_n$ are independent if

$$P(A_1 \cap \cdots \cap A_n) = P(A_1) \times P(A_2) \times \cdots \times P(A_n) = \prod_{i=1}^{n} P(A_i)$$

Definition: Random variables $Y_1 \cdots Y_n$ are independent if

$$Pr(Y_1 \in B_1 \cap \cdots \cap Y_n \in B_n) = \prod_{i=1}^{n} Pr(Y_i \in B_i)$$

Multivariate change of variables

Let $Y = (Y_1 \cdots Y_p)$ be a p-variate random variable with sample space $\mathbb{R}$. Let $X = (X_1 \cdots X_p) = (g_1(Y), \cdots g_p(Y)) = g(Y)$.

If $g(.)$ is invertible AND differentiable, then

$$p_x(x) = p_y(g^{-1}(x)) \times \mid \frac{\partial g^{-1}(x)}{\partial x} \mid$$

where the final term is the *Jacobian determinant.*

$$\begin{aligned}\boldsymbol{J} &= \left[\frac{\partial f}{\partial x_1} \cdots \frac{\partial f}{\partial x_n}\right] \\ &= \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix}\end{aligned}$$

For non-invertible functions, try to find the CDF of X from the CDFs of Y and then differentiate. Alternatively, write X as a multivariate transformation of an invertible function and then apply the multivariate change of variables formula.

Moments

Let $p(y)$ be the odf for a discrete random variable Y,

Definition: The expectation of a discrete random variable is defined as, $\mathbb{E}[Y] = \sum_{y \in \mathbb{Y}} yp(y)$

In the case of a continuous random variable, the expectation is defined as,

$$\mathbb{E}[Y] = \int yp(y)dy$$

For example, the expectation of a random variable following a beta distribution is,

$$\begin{aligned}
\mathbb{E}[y] &= \int y \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1}dy \\
&= \int \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha}(1-y)^{\beta-1}dy \\
&= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int y^{\alpha}(1-y)^{\beta-1} \frac{\Gamma(\alpha+1)\Gamma(\beta)}{\Gamma(\alpha+\beta+1)} \frac{\Gamma(\alpha+\beta+1)}{\Gamma(\alpha+1)\Gamma(\beta)} dy \\
&= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha+1)\Gamma(\beta)}{\Gamma(\alpha+1+\beta)} \\
&= \frac{\Gamma(\alpha+1)}{\Gamma(\alpha)} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha+\beta+1)} \\
&= \frac{\alpha}{\alpha+\beta}
\end{aligned}$$

Expectation of a function

Let Y be a random variable and $X = f(Y)$.

Then, $\mathbb{E}[X] = \mathbb{E}[f(Y)]$, when it exists is $\sum_{y} f(y)p_y(y)$ and $\int f(y)p(y)dy$ for discrete and continuous cases respectively.

Properties of $\mathbb{E}[.]$

- The expectation of a variable or a function is 'linear'
  Theorem: Let Y be a random variable with $Y \geq 0$ and $\mathbb{E}[Y] < \infty$. Let $a, b \in \mathbb{R}$ Then,
  $$\mathbb{E}[a + bY] = a + b\,\mathbb{E}[Y]$$
- Expectation of a joint distribution of variables is defined as,
  Let $Y_1, \cdots, Y_n$ have joint pdf $p(y_1 \cdots y_n)$
  $$\mathbb{E}[f(Y_1, \cdots Y_n)] = \int \cdots \int f(y_i \cdots y_n)p(y_1 \cdots y_n)dy_1 \cdots dy_n$$

for continuous variables and

$$\mathbb{E}[f(Y_1,\cdots Y_n)] = \sum f(y_1 \cdots y_n)p(y_1 \cdots y_n)$$

for discrete variables.

- **Theorem:**
$$\mathbb{E}[\sum a_i Y_i] = \sum a_i \, \mathbb{E}[Y_i]$$

- **Tower property**
$$\mathbb{E}[Y] = \mathbb{E}[E[Y \mid X]]$$

- **"Taking out what is known"**
$$\begin{aligned} \mathbb{E}[h(X)Y \mid X] &= h(X)\, \mathbb{E}[Y \mid X] \\ \mathbb{E}[h(X)Y \mid X = x] &= h(x)\, \mathbb{E}[Y \mid X = x] \end{aligned}$$

Variance

While the expectation is the center of mass, the variance is a measure of the spread around the center of mass.

Definition: For a random variable such that $\mathbb{E}[Y] < 0$,

$$\mathbb{V}[Y] = \mathbb{E}[(Y - \mathbb{E}[Y])^2]$$

Results that follow are,

- $\mathbb{V}[Y] = \mathbb{E}[Y^2] - \mathbb{E}[Y]^2$ if $\mathbb{E}[\mid Y \mid] < \infty$
- $\mathbb{V}[a + bY] = b^2 V(Y)$

The proof of both follow from the linearity of expectation theorem.

Proof

$$\begin{aligned} \mathbb{V}[Y] &= \mathbb{E}[(Y - \mathbb{E}[Y])^2] \\ &= \mathbb{E}[(Y^2 - 2Y\, \mathbb{E}[Y] + \mathbb{E}[Y]^2)] \\ &= \mathbb{E}[Y^2] - 2\, \mathbb{E}[Y]^2 + \mathbb{E}[Y]^2 \\ &= \mathbb{E}[Y^2] - \mathbb{E}[Y]^2 \end{aligned}$$

Proof

Define $X = a + bY$

$$\begin{aligned}
\mathbb{V}[X] &= \mathbb{V}[a + bY] \\
&= \mathbb{E}[(a + bY)^2] - \mathbb{E}[(a + bY)]^2 \\
&= \mathbb{E}[a^2 + 2abY + b^2Y^2] - (a + b\,\mathbb{E}[Y])^2 \\
&= a^2 + 2ab\,\mathbb{E}[Y] + b^2\,\mathbb{E}[Y^2] - (a^2 + 2ab\,\mathbb{E}[Y] + b^2\,\mathbb{E}[Y]^2) \\
&= a^2 + 2ab\,\mathbb{E}[Y] + b^2\,\mathbb{E}[Y^2] - a^2 - 2ab\,\mathbb{E}[Y] - b^2\,\mathbb{E}[Y]^2) \\
&= b^2\,\mathbb{E}[Y^2] - b^2\,\mathbb{E}[Y]^2 \\
&= b^2(\mathbb{E}[Y^2] - \mathbb{E}[Y]^2) \\
&= b^2\,\mathbb{V}[Y]
\end{aligned}$$

Jenson's inequality

Theorem: Suppose $f : \mathbb{R} \to \mathbb{R}$ is **convex** and $\mathbb{E}[\mid Y \mid] < \infty$,

$$\mathbb{E}[f(Y)] \geq f(\mathbb{E}[Y])$$

L_p **norm**

$$\mathbb{E}[\mid Y \mid^q]^{1/q} \leq \mathbb{E}[\mid Y \mid^p]^{1/p}$$

if $q \leq p$

Covariance

Let $Y_1 \cdots Y_n$ be random variables. The covariance of Y_i, Y_j,

$$Cov(Y_i, Y_j) = \mathbb{E}[((Y_i) - \mathbb{E}(Y_i)) \times ((Y_j) - \mathbb{E}(Y_j))]$$

For a multivariate distribution,

$$\begin{aligned}
\mathbb{V}[\textstyle\sum b_iY_i] &= \sum_i \sum_j b_ib_jCov(Y_i, Y_j) \\
&= \sum_i b_i^2\,\mathbb{V}(Y_i) + 2\sum_{i<j}\sum b_ib_jCov(Y_i, Y_j)
\end{aligned}$$

Note that if Y_i's are independent then the second term just disappears since the covariance between two independent variables is zero.

Cauchy-Schwarz inequality

Theorem: For any two random variables,

$$\mid \mathbb{E}[XY] \mid \leq \mathbb{E}[\mid XY \mid] \leq \sqrt{\mathbb{E}[X^2]\, \mathbb{E}[Y^2]}$$

This is derived from the definition of correlation- correlation between two random variables is computed as $\frac{\text{Covariance}}{\sqrt{\text{Var}_1 \text{Var}_2}}$

We know $-1 \leq Cor(Y_1, Y_2) \leq 1$. Using this, suppose $\mathbb{E}[Y_1] = \mathbb{E}[Y_2] = 0$

$$-1 \leq Cor(Y_1, Y_2) = \frac{\mathbb{E}[Y_1, Y_2]}{\sqrt{\mathbb{E}[Y_1^2]\, \mathbb{E}[Y_2^2]]}} \leq 1$$

More generally,

Holder Theorem:

$$\mid \mathbb{E}[XY] \mid \leq \mathbb{E}[\mid XY \mid] \leq \mathbb{E}[\mid X \mid^p]^{1/p} \times \mathbb{E}[\mid Y \mid^q]^{1/q}$$

if $1/q + 1/p = 1$.

Conditional expectation

Definition: The conditional expectation of Y given X is

$$\mathbb{E}[Y \mid X] = \int yp(y \mid X)dy$$

Instead of using the marginal distribution as we would for the expectation of a random variable, we use the conditional distribution. Here X is random. More generally,

$$\mathbb{E}[Y \mid X = x] = \int yp(y \mid X = x)dy$$

Inference with Sample Mean

Based on a random sample from a population, we would like to estimate the mean of a variable, describe the precision of the estimate and obtain a range of plausible values for the mean.

Sample mean estimator

To estimate the mean, we have a variety of estimators, one of which is the sample mean estimator. Let $\mathrm{Y}_1, \cdots, \mathrm{Y}_\mathrm{n} \sim$ IID P from some population P with $\mathbb{E}[\mathrm{Y_i}] = \mu, \mathbb{V}[\mathrm{Y_i}] = \sigma^2$

The sample mean estimator is defined as $\hat{\mu} = \bar{Y} = 1/n \sum Y_i$

$$\mathbb{E}[\hat{\mu}] = \mathbb{E}[1/n \sum Y_i] = 1/n \sum \mathbb{E}[Y_i] = 1/n(\sum \mu) = \mu$$

The sample mean is an unbiased estimator of μ. Note, no independence assumption was required to arrive at this conclusion.

Definition: Bias Let θ be an unknown population quantity and let $\hat{\theta}$ be a random variable. The bias of $\hat{\theta}$ for estimating θ is

$$\mathbb{E}[\hat{\theta} - \theta] = \mathbb{E}[\hat{\theta}] - \theta$$

Therefore, if the difference between the population parameter and the estimator is zero, we say the estimator is unbiased. More formally, If $\mathbb{E}_p[\hat{\theta}] - \theta(P)$ is zero for each $P \in \mathbb{P}$, then $\hat{\theta}$ is an unbiased estimator for θ in model $\mathbb{P}$. Note that both θ and $\mathbb{E}[\hat{\theta}]$ depend on the population P.

Linear Shrinkage Estimator

In statistics, shrinkage is the reduction in the effects of sampling variation. A shrinkage estimator is an estimator that either explicitly or implicitly incorporates the effects of shrinkage. In loose terms, this means that a naive or raw estimate is improved by combining it with other information. The terms relates to the notion that the improved estimator is made closer to the value supplied by the ‘other information’ than the raw estimate.

Let $\hat{\mu} = a + b\bar{Y} = (1 - w)\mu_0 + w\bar{Y}$

This is an example of a shrinkage estimator. Bayesian estimators often take this form.

Precision of $\hat{\theta}$

To quantify how close $\hat{\theta}$ is to θ, we use mean squared error

$$\text{MSE}(\hat{\theta}, \theta) = \mathbb{E}[(\hat{\theta} - \theta)^2]$$

From the definition, it follows that $\text{MSE}(\hat{\theta}, \theta) = B^2(\hat{\theta}, \theta) + \mathbb{V}[\hat{\theta}]$

Proof

This proof uses the trick of adding and subtracting $\mathbb{E}[\hat{\theta}]$ from the MSE equation

$$\begin{aligned}
\text{MSE}(\hat{\theta}, \theta) &= \mathbb{E}[(\hat{\theta} - \theta)^2] \\
&= \mathbb{E}[(\hat{\theta} - \mathbb{E}[\hat{\theta}] + \mathbb{E}[\hat{\theta}] - \theta)^2] \\
&= \mathbb{E}[((\hat{\theta} - \mathbb{E}[\hat{\theta}]) + (\mathbb{E}[\hat{\theta} - \theta]))^2] \\
&= \mathbb{E}[(\hat{\theta} - \mathbb{E}[\hat{\theta}])^2 + 2(\hat{\theta} - \mathbb{E}[\hat{\theta}])(\mathbb{E}[\hat{\theta} - \theta]) + (\mathbb{E}[\hat{\theta} - \theta])^2] \\
&= \mathbb{E}[(\hat{\theta} - \mathbb{E}[\hat{\theta}])^2] + 2(\mathbb{E}[\hat{\theta}] - \mathbb{E}[\hat{\theta}])(\mathbb{E}[\hat{\theta} - \theta]) + (\mathbb{E}[\hat{\theta} - \theta])^2 \\
&= \mathbb{V}[\hat{\theta}] + 0 + B^2(\hat{\theta}, \theta) \\
&= \mathbb{V}[\hat{\theta}] + B^2(\hat{\theta}, \theta)
\end{aligned}$$

Therefore, if bias is 0, then $\text{MSE}(\hat{\theta}, \theta) = \mathbb{V}[\hat{\theta}]$

Include MSE graph for linear shrinkage estimator and sample mean estimator

Variance of sample mean

Let $\mathrm{Y}_1, \cdots, \mathrm{Y}_\mathrm{n} \sim$ IID P from some population P with $\mathbb{E}[\mathrm{Y}_\mathrm{i}] = \mu, \mathbb{V}[\mathrm{Y}_\mathrm{i}] = \sigma^2$

$$\begin{aligned}
\mathbb{V}[\bar{Y}] &= \mathbb{V}[1/n \sum Y_i] \\
&= 1/n^2\, \mathbb{V}[\sum Y_i] \\
&= 1/n^2 \sum \mathbb{V}[Y_i] \\
&= 1/n^2 \sum \sigma^2 \\
&= \frac{\sigma^2}{n}
\end{aligned}$$

Important result! This means that $\mathbb{V}[Y_i] > \mathbb{V}[\bar{Y}]$ if $n > 1$. Further, $\mathbb{V}[\bar{Y}]$ is decreasing in n. This is the underpinnnings of all of statistics: as n increases, the distribution of $\bar{Y}$ becomes more concentrated around μ.

Confidence intervals

We want to come up with an interval of plausible values that has enough coverage to include the population parameter as well be precise so that the interval width is not $(-\infty, \infty)$.

Mathematically, this means we want $Pr(\mu \in C(Y)) = Pr(l(Y) < \mu < u(Y))$ to be close to 1 but the expected interval width $\mathbb{E}[u(Y) - l(Y)]$ to be as small as possible.

Definition: If $Pr(l(Y) \leq \mu \leq u(Y)) \geq 1 - \alpha$ for all values of μ for all populations under consideration, then $u(Y) - l(Y)$ is a $(1 - \alpha) \times 100\%$ confidence interval for μ.

Markov inequality

Let $\mathrm{Y_1, \cdots, Y_n} \sim$ IID P from some population P with $\mathbb{E}[\mathrm{Y_i}] = \mu, \mathbb{V}[\mathrm{Y_i}] = \sigma^2$

Consider an interval $C(Y) = (\bar{Y} - c, \bar{Y} + c)$. The goal is to choose a c such that $Pr(\mu \in C(Y)) \geq 1 - \alpha$.

$$\begin{aligned} Pr(\mu \in C(Y)) &= Pr(\bar{Y} - c < \mu < \bar{Y} + c) \\ &= Pr(-c < \bar{Y} - \mu < c) \\ &= Pr(\mid \bar{Y} - \mu \mid < c) \\ \text{Choose c such that,} & \\ Pr(\mid \bar{Y} - \mu \mid < c) &\geq 1 - \alpha \\ 1 - Pr(\mid \bar{Y} - \mu \mid < c) &\leq 1 - (1 - \alpha) \\ Pr(\mid \bar{Y} - \mu \mid > c) &\leq \alpha \end{aligned}$$

We want to choose a c such that the LHS → 1 as $c \to 0$ or LHS → 0 as $c \to \infty$. The interval width is 2C. The second goal is to find the smallest c i.e. smallest interval width possible.

Definition: Markov's inequality

Let X be a non-negative random variable. Then,

$$Pr(X > c) \leq \mathbb{E}[X]/c$$

Proof

$$\begin{aligned} \mathbb{E}[X] &= \int_0^\infty xp(x)dx \\ &= \int_0^c xp(x)dx + \int_c^\infty xp(x)dx \\ &\geq \int_c^\infty xp(x)dx \\ &\geq c\int_c^\infty p(x)dx \\ &= cPr(X > c) \\ \mathbb{E}[X] &\geq cPr(X > c) \\ Pr(X > c) &\leq \frac{\mathbb{E}[X]}{c} \end{aligned}$$

Deriving a confidence interval from Markov's inequality

Note, X has to be a non-negative random variable. $\mid \bar{Y} - \mu \mid$ is a non-negative random variable.

$$\begin{aligned} Pr(\mid \bar{Y} - \mu \mid > c) &\leq \mathbb{E}[\mid \bar{Y} - \mu \mid]/c \\ &\leq \mathbb{E}[\mid \bar{Y} - \mu \mid^2]^{1/2}/c = \frac{\sigma}{\sqrt{n}} \frac{1}{c} \end{aligned}$$

(This used L_p norm inequality!)

Therefore,

$$Pr(\mid \bar{Y} - \mu \mid > c) \leq \frac{\sigma}{\sqrt{n}} \frac{1}{c}$$

We want to guarantee $Pr(\mid \bar{Y} - \mu \mid > c) < \alpha$ so choose c such that,

$$\frac{\sigma}{\sqrt{n}} \frac{1}{c} \leq \alpha \Rightarrow c \geq \frac{\sigma}{\sqrt{n}} \frac{1}{\alpha}$$

Confidence interval via Markov's inequality

A $1 - \alpha$ Markov confidence interval will take the form,

$$C_m(\bar{Y}) = \left(\bar{Y} - \frac{\sigma}{\sqrt{n}} \frac{1}{\alpha}, \bar{Y} + \frac{\sigma}{\sqrt{n}} \frac{1}{\alpha} \right)$$

The Markov confidence interval used two inequalities- Markov inequality and L_p norm.

Chebyshev's inequality

Let $\mathrm{Y_1, \cdots, Y_n} \sim$ IID P from some population P with $\mathbb{E}[\mathrm{Y_i}] = \mu, \mathbb{V}[\mathrm{Y_i}] = \sigma^2$

Then,

$$Pr(\mid Y - \mu \mid > c) \leq \sigma^2/c^2$$

Proof

$$\begin{aligned} Pr(\mid Y - \mu \mid > c) &= Pr(\mid Y - \mu \mid^2 > c^2) \\ &\leq \mathbb{E}[\mid Y - \mu \mid^2]/c^2 \\ &= \sigma^2/c^2 \end{aligned}$$

We used Markov's inequality to prove Chebyshev's inequality.

In application, \ Let $\mathrm{Y_1, \cdots, Y_n} \sim$ IID P from some population P with $\mathbb{E}[\mathrm{Y_i}] = \mu, \mathbb{V}[\mathrm{Y_i}] = \sigma^2$ then,

$$Pr(\mid \bar{Y} - \mu \mid > c) \leq \frac{\sigma^2}{n} \frac{1}{c^2}$$

Confidence interval via Chebyshev's inequality,

$$\begin{aligned} Pr(\mu \in C_c(\bar{Y})) = Pr(\mid \bar{Y} - \mu \mid < c) > 1 - \alpha \\ Pr(\mid \bar{Y} - \mu \mid > c) < \alpha \\ Pr(\mid \bar{Y} - \mu \mid > c) \leq \frac{\sigma^2}{nc^2} < \alpha \end{aligned}$$

Similar to Markov's interval, here we have guaranteed $1 - \alpha$ coverage if,

$$\frac{\sigma^2}{n} \frac{1}{c^2} \leq \alpha \Rightarrow c \geq \frac{\sigma}{\sqrt{n}} \frac{1}{\sqrt{\alpha}}$$

A $1 - \alpha$ Chebyshev confidence interval will take the form,

$$C_c(\bar{Y}) = \left(\bar{Y} - \frac{\sigma}{\sqrt{n}} \frac{1}{\sqrt{\alpha}}, \bar{Y} + \frac{\sigma}{\sqrt{n}} \frac{1}{\sqrt{\alpha}} \right)$$

Normal confidence interval

When we have a normal random variable or can assume normality,

$$C_z(\bar{Y}) = \left(\bar{Y} - \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2}, \bar{Y} + \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2} \right)$$

In summary, the generic form of confidence intervals is $C(\bar{Y}) = \bar{Y} \pm \frac{\sigma}{\sqrt{n}} a$ where,

$$\begin{aligned} a_m &= \frac{1}{\alpha} \\ a_c &= \frac{1}{\sqrt{\alpha}} \\ a_z &= z_{1-\alpha/2} \\ &\text{Generally,} \\ a_m &> a_c > a_z \end{aligned}$$

Convergence in Probability

Definition 1: Let $X_1, X_2, \cdots, X_n$ be an infinite sequence of random variables. Then, $X_n \to 0$ in probability if $Pr(\mid X_n \mid > \epsilon) \to 0$ as $n \to \infty$

Definition 2: (More general case)

Let X be a random variable or a constant. Then, $X_n \to X$ in probability if $(Pr \mid X_n - X \mid > \epsilon) \to 0$ as $n \to \infty$.

So, $X_n \to X$ (X_n converges to X in probability) if $(X_n - X) \to 0$ in probability.

Here, each X has its own probability distribution- it is not representing a data point, it is a random variable with a probability distribution. The variance of the probability distribution depends on n, therefore as n increases, the variance reduces and as $n \to \infty$ the variance of X_n becomes so small that the probability of X_n converges to X. In the special case where X = 0, the probability of X_n converges to zero given $Pr(\mid X_n \mid > \epsilon) \to 0$ as $n \to \infty$.

Law of Large Numbers

Theorem: Weak Law of Large Numbers

Let $Y_1, Y_2, \cdots, Y_n \sim IID$, with $\mathbb{E}[Y_i] = \mu$, $V[Y_i] = \sigma^2 < \infty$

For each n, let $\bar{Y_n} = \frac{1}{n} \sum Y_i$. Then, $\bar{Y_n} \to \mu$ converges in probability as $n \to \infty$.

Proof

Using Chebyshev's inequality $(Pr(\mid X - \mu \mid > c) \leq \sigma^2/c)$, we get,

$$Pr(\mid \bar{Y_n} - \mu \mid > \epsilon) \leq \frac{\sigma^2}{\epsilon n} \to 0, \text{ as } n \to \infty$$

,

Using the fact that $V[\bar{Y}] = \sigma^2/n$. This is only true if the Y_is are uncorrelated.

Here, as $n \to \infty$, the quantity on the right of the inequality tends becomes 0, therefore, we get $Pr(\mid \bar{Y_n} - \mu \mid > \epsilon) \leq 0$. Using Definition 1, $\bar{Y_n} \to \mu$ in probability.

Consistency

Definition 3: For each $n \in N$, let $\hat{\theta_n}$ be an estimator of θ. Then we say $\hat{\theta}_n$ is consistent if

$$\hat{\theta}_n \to \theta, \text{ as } n \to \infty$$

in probability.

Central Limit Theorem

$$Y_1, \cdots, Y_n \sim IID$$

with $\mathbb{E}[Y_i] = \mu$, $V[Y_i] = \sigma^2 < \infty$

We can show that $\mathbb{E}[\bar{Y}] = \mu$ and $V[\bar{Y}] = \sigma^2/n$ and $\bar{Y} \to \mu$ in probability.

The central limit theorem adds to this by sayuing

$$\bar{Y} \sim N(\mu, \sigma^2/n)$$

Then,

$$\frac{\sqrt{n}}{\sigma}(\bar{Y} - \mu) \sim N(0,1)$$

$$Pr(\frac{\sqrt{n}}{\sigma}(\bar{Y} - \mu) < c) \approx \phi(c)$$

$$\lim_{n \to \infty} Pr(\frac{\sqrt{n}}{\sigma}(\bar{Y} - \mu) < c) = \phi(c)$$

This was all convergence in probability. Next, we will go through convergence in distribution.

Equality in Distribution

Definition: If $P(X_0 \in A) = P(X_1 \in A)$, we say, $X_0 = X_1$ in distribution.

Theorem: $X_0 = X_1 \Leftrightarrow F_{X_0}(a) = F_{X_1}(a)$. So, the CDFs characterize the distribution. When we say two quantities converge in distribution, we say they are converging as CDFs, not PDFs.

Theorem: $\mathbb{E}[X_0^k] = \mathbb{E}[X_1^k]$

Almost a result - $\mathbb{E}[X_0^k] = \mathbb{E}[X_1^k] \Rightarrow X_0 = X_1$ in distribution. This result only holds under some conditions, for example, if X_0, X_1 have bounded support. So in many cases, the moments of a distribution characterize the distribution.

We can check if two distributions are equal by checking the moments. However, that is alot of moments, therefore, instead we use moment generating functions to check.

Moment Generating Function

$$M_x(t) = \mathbb{E}[e^{tx}]$$

Example

Let $z \sim N(0,1)$

$$
\begin{aligned}
M_z[t] &= \mathbb{E}[e^{tz}] = \int e^{tz} p(z) dz \\
&= e^{tz} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz \\
&\text{Adding and subtracting } e^{t^2/2} \\
&= \int e^{tz} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}+\frac{t^2}{2}} e^{-\frac{t^2}{2}} dz \\
&= e^{t^2/2} \int \frac{1}{\sqrt{2\pi}} e^{-1/2(z-t)^2} dz \\
&= e^{t^2/2}
\end{aligned}
$$

Now we know that the moment generating function for a standard normal distribution is $M_z[t] = e^{t^2/2}$. How can we check moments using this function?

$$
\begin{aligned}
\frac{\partial}{\partial \mathrm{t}} M_x(t) &= \frac{\partial}{\partial \mathrm{t}} \int e^{tz} p(x) dx \\
&= \int \frac{\partial}{\partial \mathrm{t}} e^{tx} p(x) dx \\
&= \int x e^{tx} p(x) dx \\
\text{So, } \frac{\partial}{\partial \mathrm{t}} M_t(t)_{t=0} &= \int x p(x) dx = \mathbb{E}[x]
\end{aligned}
$$

The number of times we differentiate the moment generating function is the order of the moment we generate! The moment generating functionis always evaluated at t = 0.

$$
\frac{\partial^{\mathrm{k}}}{\partial \mathrm{t}^{\mathrm{k}}} M_x(t)_{t=0} = \mathbb{E}[x^k]
$$

In summary, the moment generating functions tell us about the moments and then, the moments tell us about the distribution.

Theorem: If $M_{x_0}(t)$ and $M_{x_1}(t)$ exists and $M_{x_0}(t) = M_{x_1}(t)$, then,

$$
X_0 = X_1
$$

.

Therefore, instead of checking the each individual moment in order to find whether two variables are equal in distribution, we check to see if the moment generating functions are equal to each other. If the moment generating functions exist and are equal, then we can say that the two random variables are equal in distribution. However, if the moment generating functions do not exist, then

$$
\mathbb{E}[X_0^k] = \mathbb{E}[X_1^k]
$$

but $X_0 \neq X_1$. And then, there are cases when $M_{X_1}(t) \approx M_{X_0}(t)$ but distributions of X_0, X_1 are quite different.

Convergence in Distribution

From the last section we have, **Definition 3:** $X_1 = X_0$ if $F_{X_0}(a) = F_{X_1}(a)$.

For finding *convergence* in probability, we have

Definition 4: $X_n \to X_0$ if $F_{x_n}(x) \to F_{x_0}$. (Remark: We only need to check at points of continuity).

Example If $Pr(\frac{\sigma}{\sqrt{n}} \mid \bar{Y} - \mu \mid > c) \to \phi(c)$, then, $\left(\frac{\sigma}{\sqrt{n}}\bar{Y} - \mu\right) \to z$ in distribution.

Theorem: Convergences of MGFs

Let $X_1, X_2, \cdots,$ be a sequence of random variables with MGFs $MGF_{x_1}, MGF_{x_2}, \cdots$. If, $M_{x_n}(t) \to M_{x_0}(t)$ as $n \to \infty$, Then, $F_{x_n}(t) \to F_{x_0}(t)$ as $n \to \infty$ which also means $X_n \to X_0$ i.e. X_n converges to X_0 in distribution.

Proof Prove $Pr(\bar{z}_n < c) \to Pr(z < c) = \phi(c)$.

In order to prove that these two distributions are converging, we can show that the two moment generating functions are converging i.e. $M_{z_n}(t) \to M_z(t)$.

First we find the MGF of $\bar{z_n}$

$$\begin{aligned}
\bar{z_n} &= \frac{\sqrt{n}}{\sigma}\left(\sum Y_i - \mu\right) = \frac{1}{\sqrt{n}}\sum\left(\frac{Y_i - \mu}{\sigma}\right) \\
&= \frac{1}{\sqrt{n}}\sum z_i \\
M_{\bar{z_n}}(t) &= \mathbb{E}[e^{t\bar{z}}] = \mathbb{E}[e^{tz_i}]^n
\end{aligned}$$

To find $M_{z_i}(t)$ we use the Taylor approximation series, In mathematics, a Taylor series is a representation of a function as an infinite sum of terms tha tare calculated from the values of the function's deriviates at a single point. A function can be approximated by using a finite number of terms of it's Taylor series. The Taylor series of a real or complex-valued function that is infinitely differentiable at a real or complex number a is the power series,

$$f(x) = f(a) + \frac{f'(a)}{1!} + \frac{f''(a)}{2!}(x-a)^2 + \frac{f'''(a)}{3!}(x-a)^3 + \cdots$$

In compact σ notation, this can be written as,

$$\sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!}(x-a)^n$$

where f^n is the nth derivative of f evaluated at point a.

So,

Define s $= \dfrac{\sqrt{t}}{n}$

$$M_{\bar{z_n}} = M_{z_i}(\frac{t}{\sqrt{n}}))^n$$

$$M_{z_i}(s) = sM_{z_i}(0) + sM'_{z_i}(0) + \frac{s^2}{2}M''_{z_i}(0) + \frac{s^3}{3!}M'''_{z_i}(0) + \cdots$$

Note, here z = 0

$$M_{z_i}(s) = s\,\mathbb{E}[e^{sz}] + s\,\mathbb{E}[z_i] + \frac{s^2}{2}\,\mathbb{E}[z_i^2] + \cdots$$

$$M_{z_i}(t/\sqrt{n}) = 1 + 0 + \frac{t^2/2}{n}$$

Therefore,

$$M_{\bar{z_n}} = (1 + \frac{t^2/2}{n})^n$$

Using $\log(1 + a/n)^n = n\ \log(1 + a/n) \approx a$ for $n \to \infty$

$$M_{\bar{z_n}} = e^{t^2/2}$$

which is the moment generating function for a standard normal distribution. Therefore, for $n \to \infty, \bar{z_n} \to z$ in distribution.

Consistent and Asymptotically Normal (CAN) Estimators

In the previous sections, we have shown that if $Y_1, \cdots, Y_n \sim \ IID, \mathbb{E}[Y_i] = \mu, V[Y_i] = \sigma^2$,

Then,

a. $\bar{Y} \to \mu$ in probability
b. $\sqrt{n}(\bar{Y} - \mu)/\sigma \to N(0,1)$ in distribution

Definition 1: Let $\hat{\theta}$ be a sequence of estimators such tat,

$$\frac{\sqrt{n}}{\tau}(\hat{\theta} - \theta) \to N(0,1)$$

Then, $\hat{\theta}$ is CAN for θ.

Asymptotic Confidence Intervals

Let $z_{1-\alpha/2}$ be such that $Pr(Z > z_{1-\alpha/2} = \alpha/2$. Consider the random interval, $C(\hat{\theta}) = \hat{\theta} \pm \tau/\sqrt{n}z_{1-\alpha/2}$. Then,

$$\begin{aligned}
Pr(\theta \in C(\hat{\theta})) &= Pr(\theta - \tau/\sqrt{n}z_{1-\alpha/2} < \hat{\theta} < \theta + \tau/\sqrt{n}z_{1-\alpha/2}) \\
&= Pr(z_{1-\alpha/2} < \frac{\sqrt{n}}{\tau}(\hat{\theta} - \theta) < z_{1-\alpha/2}) \\
&\approx \phi(z_{1-\alpha/2}) - \phi(-z_{1-\alpha/2}) \quad (CAN) \\
&= 1 - \alpha/2 - (\alpha/2) \\
&= 1 - \alpha
\end{aligned}$$

Continuous Mapping Theorem: If $\hat{\theta} \to \theta$ in probability, then,

$$\hat{\psi_n} \equiv g(\hat{\theta}_n) \to g(\theta) \equiv \psi$$

(in probability).

Proof

We need to show that $\hat{\psi}_n \to \psi$ so we prove $Pr(\mid \hat{\psi} - \psi \mid > \epsilon) \to 0$.

Choose ϵ and δ such that,

$$\mid \hat{\theta_n} - \theta \mid < \delta \Rightarrow \mid g(\hat{\theta_n}) - g(\theta) \mid < \epsilon$$

Conversely,

$$\mid g(\hat{\theta_n}) - g(\theta) \mid > \epsilon \Rightarrow \mid \hat{\theta_n} - \theta \mid > \delta$$

Therefore,

$$Pr(\mid g(\hat{\theta}_n) - g(\theta) \mid > \epsilon) < Pr(\mid \hat{\theta}_n - \theta \mid > \delta) \to 0$$

as $n \to \infty$.

So, $\hat{\psi} \equiv g(\hat{\theta}_n)$ is a consistent estimator for $\psi \equiv g(\theta)$.

Implication: A consistent estimator for θ provides a consistent estiamtor for $g(\theta)$ for any continuous g!

Delta method

Constructing a confidence interval...

Let $\hat{\theta_n}$ be CAN for θ, then we can show $\sqrt{n}(\hat{\theta} - \theta)/\tau \to N(0, 1)$.

Define $\hat{\psi_n} = g(\hat{\theta_n})$. Is $\hat{\psi_n}$ CAN for ψ?

Theorem: If g is differentiable such that $g'(\theta) \neq 0$, then,

$$\frac{\sqrt{n}(\hat{\psi} - \psi)}{g'(\theta)\tau} \to N(0, 1)$$

So, $\hat{\psi_n}$ is CAN for ψ,

$$\hat{\psi} \sim N(\psi, \left(g'(\theta)^2\tau^2\right)/n)$$

To prove this, we use the Mean Value Theorem.

Proof

$$\hat{\psi_n} = g(\hat{\theta_n}) \approx g(\theta) + g'(\theta)(\hat{\theta_n} - \theta)$$

$$\frac{\hat{\psi_n} - \psi}{g'(\theta)} \approx \hat{\theta_n} - \theta$$

Multiplying both sides by square root n and τ,

$$\frac{\sqrt{n}\hat{\psi_n} - \psi}{g'(\theta)\tau} \approx \frac{\sqrt{n}\hat{\theta_n} - \theta}{\tau} \to N(0, 1)$$

Convergence in probability is convergence in distribution

Proof To prove that convergence in probability is convergence in distribution, we use the squeeze theorem.

Given $Pr(\mid X_n - X \mid > \epsilon) \to 0$, we want to show $Pr(X_n \leq x) \to Pr(X \leq x)$

$$\begin{aligned}
Pr(X_n \leq x) &= Pr(X_n \leq x, X \leq x+\epsilon) + Pr(X_n \leq x, X > x+\epsilon) \\
&\leq Pr(X \leq x+\epsilon) + Pr(\mid X_n - X \mid > \epsilon) \\
&\lim_{n\to\infty} Pr(Xn \leq x) \leq Pr(X \leq x+\epsilon) \\
&Pr(X \leq x-\epsilon) \leq \lim_{n\to\infty} Pr(X_n \leq x) \\
Pr(X \leq x-\epsilon) &= \lim_{n\to\infty} Pr(X_n \leq x) \leq Pr(X \leq x+\epsilon) \\
&= F(x-\epsilon) \leq \lim_{n\to\infty} F_n(x) \leq F(x+\epsilon) \\
&\text{Taking a limit of } \epsilon \to 0 \\
&= F(x) \leq \lim_{n\to\infty} F_n(x) \leq F(x) \\
&\text{so, } \lim_{n\to\infty} F_n(x) = F(x)
\end{aligned}$$

Sultzky's theorem:

$$\begin{aligned}
&X_n \to X \text{ and } Y_n \to C \text{ then,} \\
&\qquad \text{A. } X_n + Y_n \to X + C \\
&\qquad \text{B. } X_n Y_n \to XC
\end{aligned}$$

Note: To prove this, use the "plug-in" approach.

Multivariate CLT and Delta Method

Theorem (Multivariate Central Limit Theorem): Suppose $\boldsymbol{Y_1, \cdots, Y_n} \sim IID$, with $\boldsymbol{Y_i} \in \mathbb{R}^p, \mathbb{E}[\boldsymbol{Y_i}] = \mu, V[\boldsymbol{Y}_i] = \Sigma$.

Then,

$$\sqrt{n}\Sigma^{-1/2}(\bar{\boldsymbol{Y}} - \boldsymbol{\mu}) \to N(0, \mathbb{I})$$

Theorem (Multivariate Delta Method): Suppose $\sqrt{n}(\bar{\boldsymbol{Y}} - \boldsymbol{\mu}) \to N(0, \Sigma)$ and, $g : \mathbb{R}^p \to \mathbb{R}$ is differentiable, with gradient,

$$\nabla g(\boldsymbol{Y}) = \begin{pmatrix} dg(\boldsymbol{y}/dy_1) \\ \vdots \\ dg(\boldsymbol{y}/dy_p) \end{pmatrix}$$

Then,

$$\sqrt{n}(g(\bar{\boldsymbol{Y_n}}) - g(\boldsymbol{\mu})) \to N(0, [\nabla g(\mu)^T \Sigma \nabla g(\mu)])$$

Parameteric Inference and Likelihood Methods

A parametric statistical model is a set of probability distributions, indexed by a finite dimensional parameter.

Consider a population P_{θ_0}. Let data come from $Y_1, \cdots, Y_n \sim iid \quad P_{\theta_0}$. Non parameteric inference proceeds by estimating P_{θ_0}, $g(P_{\theta_0})$. With parametric inference however, we estimate P_{θ_0} and $g(P_{\theta_0})$ assuming that $P_{\theta_0} \in \{P_{\theta_0} : \theta \in \Theta\} \equiv \mathbb{P}$.

Parametric inference should be easier than nonparametric inference, assuming that the model is correct.

Likelihood and MLEs

Let $\mathbb{P} = \{P_{\theta_0} : \theta \in \Theta\}$ be a model and let $f(y \mid \theta)$ be the density of P_{θ_0}. Therefore the joint density is

$$p(y \mid \theta) = \prod_{i=1}^{n} f(y_i \mid \theta)$$

Likelihood function: For data values $y_1, y_2, \cdots, y_n$ and model $\mathbb{P} = \{P_{\theta_0} : \theta \in \Theta\}$, the likelihood function is

$$L(\theta : y_1, \cdots, y_n) = \prod_{i=1}^{n} f(y_i \mid \theta) \qquad \theta \in \Theta$$

Likelihood is the density at $y_1, \cdots, y_n$ as a function of θ.

Log likelihood function:

$$\begin{aligned} l(\theta : y_1, \cdots, y_n) &= \log \mathrm{L}(\theta : y_1 \cdots y_n) \\ &= \log \ \prod f(y_i \mid \theta) \\ &= \sum \log \ f(y_i \mid \theta) \end{aligned}$$

Example Which θ values makes the data “most probable”? Normal variance example.

$$\begin{aligned} l(\theta : \boldsymbol{y}) &= -n/2[\log \theta + 1/\theta \sum y_i^2/n] \\ \frac{d}{d\theta} l(\theta : \boldsymbol{y}) &= -n/2[1/\theta - 1/\theta^2 \sum y_i^2/n] \\ -n/2[1/\theta - 1/\theta^2 \sum y_i^2/n] &= 0 \\ \hat{\theta} &= \frac{\sum y_i^2}{n} \end{aligned}$$

Maximum Likelihood Estimator: THe MLE of θ is the value of θ that maximizes the likelihood

$$\hat{\theta}_{MLE} = \text{arg max}_{\theta \in \Theta}\, L(\theta : \boldsymbol{y})$$

Example

Let $Y_1 \cdots Y_n \sim$ Unifrom $(0, \theta)$,

$$\begin{aligned} f(y_i \mid \theta) &= \frac{1}{\theta}\mathbb{I}(y_i \leq \theta) \\ L(y_i \mid \theta) &= \prod_{i=1}^{n} \frac{1}{\theta}\mathbb{I}(y_i \leq \theta) \\ &= \frac{1}{\theta^n}\prod_{i=1}^{n}\mathbb{I}(y_i \leq \theta) \\ &= \frac{1}{\theta^n}\mathbb{I}(y_{(n)} \leq \theta) \end{aligned}$$

Looking at this equation, we see that the value of θ that would maximize the likelihood is the maximum value of y_i. Therefore the MLE for the uniform distribution is $\hat{\theta} = \max(Y_1, \cdots, Y_n)$.

Consistency of MLE

Previously we've shown sample moments are consistent for population moments using law of large numbers. Now we are using LLN to show consistency of MLEs.

$$\begin{aligned} \hat{\theta}_{MLE} &= \text{arg max}_\theta\, l(\theta : \boldsymbol{y}) \\ &= \text{arg max}_\theta \sum \log f(y_i \mid \theta) \\ &= \text{arg max}_\theta\, 1/n \sum \log f(y_i \mid \theta) \end{aligned}$$

Let θ be a possible value of θ, let θ_0 be the true value and $Y \sim P_{\theta_0}$,

$$1/n \sum \log f(y_i \mid \theta) \to \mathbb{E}_{\theta_0}[\log f(Y \mid \theta)]$$

$$\begin{aligned} \mathbb{E}_{\theta_0}[\log f(Y \mid \theta)] &= \int (\log f(y \mid \theta)) f(y \mid \theta_0) dy \\ \text{for large n,} & \\ \frac{l(\theta : y)}{n} &\approx \int (\log f(y \mid \theta)) f(y \mid \theta_0) dy \\ \text{arg max}_\theta \frac{l(\theta : y)}{n} &\approx \text{arg max}_\theta (\log f(y \mid \theta)) f(y \mid \theta_0) dy \\ &\equiv \hat{\theta}_0 \end{aligned}$$

Lemma: $\hat{\theta}_0 = \theta_0$.

Proof.

$$\begin{aligned}
\mathbb{E}[\log f(Y \mid \theta)] &= \mathbb{E}[\log f(Y \mid \theta)] - \mathbb{E}[\log f(Y \mid \theta_0)] + \mathbb{E}[\log f(Y \mid \theta_0)] \\
&= \mathbb{E}[\log \frac{f(Y \mid \theta)}{f(Y \mid \theta_0)}] + c \quad \text{c is a function of } \theta \\
\mathbb{E}[\log \frac{f(Y \mid \theta)}{f(Y \mid \theta_0)}] &\leq \log[\mathbb{E}[\frac{f(Y \mid \theta)}{f(Y \mid \theta_0)}]] \\
&= \log \int \frac{f(Y \mid \theta)}{f(Y \mid \theta_0)} . f(Y \mid \theta_0) dy \\
&= \log \int f(y \mid \theta) dy \\
&= \log 1 = 0 \\
0 &\geq \mathbb{E}[\log \frac{f(Y \mid \theta)}{f(Y \mid \theta_0)}] \\
&= \mathbb{E}[\log f(Y \mid \theta) - \log f(Y \mid \theta_0)] \\
&= \mathbb{E}[\log f(Y \mid \theta)] - \mathbb{E}[\log f(Y \mid \theta_0)] \\
\mathbb{E}[\log f(Y \mid \theta_0)] &\geq \mathbb{E}[\log f(Y \mid \theta)]
\end{aligned}$$

So $\mathbb{E}[\log f(Y \mid \theta)]$ is maximized at $\theta = \theta_0$,

$$\arg\max \mathbb{E}[\log f(Y \mid \theta)] = \theta_0$$

M Estimation

For each n, let $M_n(\theta)$ be a random function or a function of θ and data or $1/n \sum \log f(y_i \mid \theta)$.

Suppose $M_n(\theta) \to M(\theta)$ for each $\theta \in \Theta$.

Theorem: Let $M_n(\theta)$ be a random function for each $n \in \mathbb{N}$. Let $M(\theta) : \Theta \to \mathbb{R}$. If,

$$\sup_{|\theta - \theta_0| > \epsilon} M(\theta)$$

and

$$\sup_{\theta \in \Theta} \mid M_n(\theta) - M(\theta) \mid \to 0$$

Then,

$$\hat{\theta} \to \theta_0$$

where, $\hat{\theta}_n = \arg\max_{\theta} M_n(\theta)$

Identifiability

Definition: A model $\mathbb{P} = \{P_\theta : \theta \in \Theta\}$ is identifiable if

$$\theta_1 \neq \theta_2 \Rightarrow P_{\theta_1} \neq P_{\theta_2}$$

Identifiabiliy is really about the parameterization of the model.

Example Let $\mathbb{P} = \{N(a+b, \sigma^2), (a, b, \sigma^2) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}^+\}$. Then, $\theta_1 = (a, b, \sigma^2)$ and $\theta_2 = (a+c, b-c, \sigma^2)$ so $\theta_1 \neq \theta_2$ but $P_{\theta_1} = P_{\theta_2}$. This model parameterization is not identifiable.

Similarly with linear regression, all the predictors have to be linearly independent otherwise the model is not identifiable.

Theorem: Let $M(\theta) : \Theta \to \mathbb{R}$ and for each $n \in \mathbb{N}$ let $M_n(\theta)$ be a random function such that,

$$\sup_{\theta : |\theta - \theta_0| > \epsilon} M(\theta) < M(\theta_0)$$

Then,

$$\sup_{\theta \in \Theta} |M_n(\theta) - M(\theta)| \to 0$$

in probability. Then, $\hat{\theta}_n \to \theta$ in probability where $\hat{\theta}_n = \arg\max_\theta M_n(\theta)$

Proof...

Insert here

Information and Asymptotic Normality

To study the properties of $\hat{\theta}_{MLE}$ is it useful to define the score function.

Definition: The score function is $S(\theta : y) = \frac{d}{d\theta} \log f(y|\theta)$

Interpretation $\Rightarrow$

$$\begin{aligned}
\frac{1}{n} l(\theta : y) &= \sum \log f(y_i|\theta)/n \\
\frac{d}{d\theta}\frac{1}{n} l(\theta : y) &= \sum S(\theta : y_i)/n \\
\frac{d}{d\theta}\frac{1}{n} \log(\theta_{MLE} : y) &= \sum S(\theta_{MLE} : y_i)/n = 0
\end{aligned}$$

The sample mean of the score function is zero at the MLE!

Fisher Information

$$\begin{aligned} I_n(\theta) &= \mathbb{E}[\hat{I_n}(\theta)] = -n\,\mathbb{E}[\frac{d^2}{d\theta^2}\log f(Y|\theta)] \\ &= -n\,\mathbb{E}[\frac{d}{d\theta}S(\theta : Y)] \\ &= n \times I(\theta_0) \end{aligned}$$

This is the "Fisher information" or the "expected information". It is the amount of information we would expect from a sample.

Insert examples of poisson model and normal model here

Variance of score information

Theorem $V[S(\theta_0 : Y)] = I(\theta_0)$.

Proof.

Asymptotic normality

Hypothesis testing

Consider two populations P_A, P_B with means μ_A, μ_B respectively. The hypothesis we want to test is whether $\mu_A = \mu_B$? We have datasets $Y_1^A, \cdots Y_n^A \sim P_A, Y_1^B, \cdots Y_n^B \sim P_B$. Let $\boldsymbol{Y} = (Y_1^A \cdots Y_n^A, Y_1^B \cdots Y_n^B)$. Based on this $\boldsymbol{Y}$, we will decided whether $\mu_A = \mu_B$.

Statistical hypothesis testing

Let θ be some unknown quantity (for example $\theta = \mu_A - \mu_B$). Our task is to evaluate the hypothesis $H : \theta = \theta_0$ where $\theta_0 = 0$ if $\mu_A = \mu_B$.

The procedure is as follows- first compute a test statistic $t(\boldsymbol{Y})$ for example, $t(\boldsymbol{Y}) = \bar{Y_A} - \bar{Y_B}$ as a function of observed data.

Then, we accept our hypothesis H for some values of $t(\boldsymbol{Y})$ of $t(\boldsymbol{Y}) \in A_{\theta_0}$ which is the "acceptance region", and reject it if not.

Often, the acceptance region takes the form $A_0 = (-a, a)$. The question arises, how big should a be?

Level α tests

- If H is true: Accept H → Correct action
- If H is true: Reject H → Type I error
- If H_1 is true: Accept H_1 → Correct action
- If H_1 is true: Reject H_1 → Type II error

If the null hypothesis H is true, we want the probability of rejection to be small (i.e. control the Type I error). Likewise, if H is false we want the probability of rejection to be big (i.e. have big power).

Formally,

$$Pr(\text{reject } H \mid H) = Pr(t(\boldsymbol{Y} \notin A \mid H)$$

→ Type I error rate

$$Pr(\text{reject } H \mid \text{not } H) = Pr(t(\boldsymbol{Y}) \notin A \mid \text{not } H)$$

→ Power

Defintion: A test (t, A) is a level-α test if $Pr(t(\boldsymbol{Y}) \in A \mid H) \leq \alpha$ i.e. the type I error rate is $\leq \alpha$.

Example: Two sample comparisons

Our test statistic is $t(\boldsymbol{Y}) = \bar{Y_A} - \bar{Y_B}$ and the acceptance region is $A = (-a, a)$.

Then, the type I error rate is $Pr(t(\boldsymbol{Y}) \notin A \mid H) = Pr(\mid \bar{Y_A} - \bar{Y_B} \mid > a \mid H)$

Using CLT, we know $\bar{Y_A} \sim N(\mu_A, \sigma_A^2/n_A)$ and $\bar{Y_B} \sim N(\mu_B, \sigma_B^2/n_B)$

Then,

$$\bar{Y_A} - \bar{Y_B} \sim N(\mu_A - \mu_b, \sigma_A^2/n_A + \sigma_B^2/n_B)$$

$$\frac{\bar{Y_A} - \bar{Y_B}}{\sqrt{\sigma_A^2/n_A + \sigma_B^2/n_B)}} \sim N(0,1)$$

if $\mu_A = \mu_B$.

Let $\sigma_d = \sqrt{\sigma_A^2/n_A + \sigma_B^2/n_B)}$

So,

$$\begin{aligned} Pr(\mid \bar{Y_A} - \bar{Y_B} \mid > \alpha \mid H) &= Pr(\frac{\mid \bar{Y_A} - \bar{Y_B} \mid}{\sigma_d} > \frac{\alpha}{\sigma_d} \mid H) \\ &\approx Pr(|z| > \frac{\alpha}{\sigma_d} \mid H) \\ &= 2 \times (1 - \phi(\frac{\alpha}{\sigma_d})) \\ &= 2 \times \phi(-\frac{\alpha}{\sigma_d}) \end{aligned}$$

this will be a level α test if

$$\begin{aligned} 2(1 - \phi(a/\sigma_d)) &\leq \alpha \\ 1 - \phi(a/\sigma_d) &\leq \alpha/2 \\ \phi(a/\sigma_d) &\geq 1 - \alpha/2 \\ a &\geq \sigma_d z_{(1-\alpha/2)} \end{aligned}$$

In summary, the approximate level α test of H : $\mu_A = \mu_B$ is,

$t(\boldsymbol{Y}) = \bar{Y_A} - \bar{Y_B}$, A $= (-z_{1-\alpha/2}\sqrt{\sigma_A^2/n_A + \sigma_B^2/n_B}, +z_{1-\alpha/2}\sqrt{\sigma_A^2/n_A + \sigma_B^2/n_B})$

The problem however is we don't usually know the population parameter σ. Therefore the approximate solution is to use sample variances $s_A^2 = \frac{1}{n_A - 1}\sum(Y_{iA} - \bar{Y_A})^2$.

Therefore,

$$t(\boldsymbol{Y}) = \frac{\bar{Y_A} - \bar{Y_B}}{\sqrt{s_A^2/n_A + s_B^2/n_B}}$$

This is called the two sample t-statisitc (with unequal variances).

Then for large n, $t(y) \sim N(0, 1)$ under the null hypothesis and if the populations are normal $t(y) \sim t_{n_A + n_B - 2}$

Note: For small n_A, n_B, one typically uses t-quantiles instead of z-quantiles for the acceptance region. This helps account for the face that the sample variances are estimates.

Hypothesis testing and CI

Defintion: $C(Y)$ is a $1 - \alpha$ CI for a parameter $\theta \in \Theta$ if,

$$Pr(\theta_0 \in C(Y) \mid \theta = \theta_0) \geq 1 - \alpha$$

Proof

$$
\begin{aligned}
Pr(\text{reject } H \mid H \text{ is true}) &= Pr(\theta_0 \notin C(Y) \mid \theta = \theta_0) \\
&= 1 - Pr(\theta_0 \in C(Y) \mid \theta = \theta_0) \\
&\leq 1 - (1 - \alpha) = \alpha
\end{aligned}
$$

Randomization and Permutation tests

In a randomization test, the random thing is the treatment of A or B. The testing procedure is as follows-

- Obtain data
- Compute a testing statistic
- Compare the test statiistic to values we would see if H were true in other words, compare the test statistic to a null distribution

Remember, under the randomization scheme, each treatement assignment is equally likely. We can randomly assign the treatment assignments over and over again to find the null distribution (the data remains fixed). ONce we have the null distribution, we compare the test statisitc to the null. The null distribution is often approximated by Monte Carlo sampling.

However, in simple experiments, the null distribution can be obtained by recomputing the test statistic under all permutations of the treatment assignment, this sort of test is called a "permutation test".

Under complex experiment designs, the permutation test is inappropriate.

P-values

Recall, we are testing the null hypothesis that H: $\theta = \theta_0$. The test statistic is $t(y)$ and acceptance region is $A(\theta_0)$.

The test is a level α test as long as

$$Pr(t(Y) \in A(\theta_0) \mid \theta = \theta_0) \leq \alpha$$

P-value: A quantification of the magnitude of $t(Y)$ relative to $t(Y^*)$ where Y^* is a random outcome under the null hypothesis.

p-value $= Pr(t(Y^*) \geq t \mid H)$ in other words, *The probability of observing data as or more extreme as the observed data if the null hypothesis were true.*

If the pvalue is small, then $t(Y)$ is extremely unlikely under the null hypothesis and therefore the null hypothesis may be inconsistent If the p-value is big, then $t(Y)$ is not extreme with respect to the null hypothesis.

In other words,

$$
\begin{aligned}
pv(t) &= Pr(|t(Y^*)| > |t| \mid H) \\
&\geq \alpha \text{ if } \mid t \mid \leq t_{1-\alpha/2} \\
&\leq \alpha \text{ if } \mid t \mid \geq t_{1-\alpha/2}
\end{aligned}
$$

Generally speaking, this means that the p-value is the smallest α at which the null hypothesis is rejected.

Theorem Let T have a continuous distribution under H. Then $Pr(pv(T) \leq \alpha) = \alpha$.

Proof

$$
\begin{aligned}
pv(t) < \alpha &= Pr(T^* \geq t \mid H) \leq \alpha \\
&= t > t_{1-\alpha} \\
Pr(pv(T) < \alpha \mid H) &= Pr(T > t_{1-\alpha} \mid H) = \alpha
\end{aligned}
$$

The p-value has a uniform distribution under the null hypothesis.

Multiple camparisons

Normal Means Model : $z_1, \cdots z_m$ iid and $z_i \sim N(\theta_i, 1)$ or $\boldsymbol{z} \sim N(\boldsymbol{\theta}, \mathbb{I})$

This model can be used for inference in a variety of situations

Multigroup effects

Goal: The goal here is to assess differences between treatment A and treatment B under a variety of conditions, or in a variety of populations.

Let $\mu_{jA} = \mathbb{E}[Y_{ijA}], \mu_{jB} = \mathbb{E}[Y_{ijB}]$ Then, $z_j = \frac{\bar{Y}_{jA} - \bar{Y}_{jB}}{s_d} \sim N(\theta_j, \mathbb{I})$

where, $\theta_j = \frac{\mu_A - \mu_B}{\sigma_j \sqrt{1/n_A + 1/n_B}}$

Linear regression

$$\begin{aligned}
Y &\sim N_n(X\beta, \sigma^2\mathbb{I}) \\
\hat{\beta}_{MLE} &= (X^TX)^{-1}X^Ty \\
\mathbb{E}[\hat{\beta}_{MLE}] &= (X^TX)^{-1}X^T\,\mathbb{E}[Y] \\
&= \beta \\
\mathbb{V}[\hat{\beta}_{MLE}] &= (X^TX)^{-1}\sigma^2 \\
\hat{\beta} &\sim N_p(\beta, (X^TX)^{-1}\sigma^2) \\
\text{Multiple by } (X^TX)^{-1/2}, & \\
\hat{\alpha} &= (X^TX)^{-1/2}\hat{\beta} \sim N((X^TX)^{1/2}\beta, \mathbb{I}\sigma^2) \\
z &= \hat{\alpha}/\hat{\sigma} \sim N_p((X^TX)^{1/2}\beta/\sigma, \mathbb{I}) \\
&= N_p(\theta, \mathbb{I})
\end{aligned}$$

And multiple testing0 continued in the next chapter

Multiple Testing

Consider testing $H_j : \theta_j = 0$ for each $j = 1 \cdots m.$ We reject the null hypothesis if $\mid z_j \mid > z_{1-\alpha/2}$

The p-value is $= Pr(\mid z \mid > \mid z_j \mid) = 2(1 - \phi(\mid z_j \mid)$.

Suppose we reject $H_j : \theta_j = 0$ if $p_j < \alpha$ for each j, then $Pr(\text{reject } H_j \mid H_j \text{ true}) = \alpha =$ type I error rate

However, what is the *global* error rate?

Global Null and Error

The global null hypothesis is : H_0 : all H_j are true or $\theta_j = 0$.

The testing procedure follows by rejecting the global null hypothesis if any H_j are rejected at level α.

The global error rate on the other hand is $Pr(\text{reject } H_0 \mid H_0 \text{ true}) =$ FWER (family wide error rate).

$$\begin{aligned}
Pr(\text{ reject } anyH_j \mid \text{ all } H_j \text{ true}) &= Pr(\text{any } p_j < \alpha \mid H_0) \\
&= 1 - Pr(\text{no } p_j < \alpha \mid H_0) \\
&= 1 - Pr(allp_j > \alpha \mid H_0) \\
&= 1 - Pr(p_1 > \alpha \cap p_2 > \alpha \cap \cdots \cap p_m > \alpha \mid H_0) \\
&= 1 - Pr(p_1 > \alpha)Pr(p_2 > \alpha) \cdots Pr(p_m > \alpha) \\
&= 1 - (1-\alpha)^m
\end{aligned}$$

Global control vida Bonferroni's correction

The goal is to come up with a global error rate below α_0. Then proceed by rejecting H_0 if any H_j is rejected at level α. We have to choose an α such that the global error rate is controlled.

If all tests are independent,

$$\begin{aligned}
Pr(\text{ reject } H_0 \mid H_0) = 1 - (1-\alpha)^m &= \alpha_0 \\
(1-\alpha)^m &= 1 - \alpha_0 \\
1 - \alpha &= (1-\alpha_0)^{1/m} \\
\alpha &= 1 - (1-\alpha_0)^{1/m}
\end{aligned}$$

Therefore, we set $\alpha = 1 - (1-\alpha_0)^{1/m}$ so that we can control the gloabl error rate.

Bonferroni's correction →

$$\begin{aligned}
\alpha &= 1 - (1-\alpha_0)^{1/m} \\
1 - \alpha &= (1-\alpha_0)^{1/m} \\
\log(1-\alpha) &= 1/m \log(1-\alpha_0) \\
-\alpha &\approx -(1/m)\alpha_0 \Rightarrow \alpha \approx \alpha_0/m
\end{aligned}$$

Bonferroni Method

Let p_j be the p-value under H_j. Let $H_0 \cap H_j$ where $H_0 =$ all H_j are true To control the global type I error rate, $Pr(\text{rej } H_0 \mid H_0) \leq \alpha_0$.

Reject H_0 if $p_j \leq \alpha_0/m$ for any $j = 1 \cdots m$.

The previous proofs showed that this will control error.

Theorem

Let $p_j \sim Unif(0,1)$ under $H_j, h = 1 \cdots m$ (not necessarily independent). Then, if we reject H_0 if any $p_j \leq \alpha_0/m$,

$$Pr(\text{ reject } H_0 \mid H_0) \leq \alpha_0$$

Proof

$$\begin{aligned} Pr(\text{ reject } H_0 \mid H_0) &= Pr(\text{ reject } H_1 \cup \text{ reject } H_2 \cup \cdots \cup \text{ reject } H_m \mid H_0) \\ &= Pr(\cup_{j=1}^{n} p_j < \alpha_0/m \mid H_0) \\ &\leq \sum Pr(p_j < \alpha_0/m \mid H_0) = \sum_j^m \alpha_0/m = \alpha_0 \end{aligned}$$

Bonferroni's method conrols the global error rate below α_0, even if tests or p-values are dependent. Therefore for example, can be used with correlated estimates of regression coefficients.

Limitations of Bonferroni

$$Pr(\text{ reject } H_0 \mid H_0 \text{ false}) = ?$$

Scenario 1: Needle in a haystack

$$\theta_k \neq 0, \theta_j = 0 \text{ for } j \neq k$$

but don't know what k is

Scenario 2: Many small effects All $\theta_j \neq 0$ but very small effects

Rejecting the null hypothesis using Bonferroni's method requires just one p-value to be small or just one effect to be large.

If the effect size is large in the first scenario then the pvalue is likely to be small therefore $Pr(pv \leq \alpha_0/m \mid \theta_k \text{ large})$ is large. Therefore Bonferroni has decent power under Scenario 1.

But under Scenario 2, if all $\theta_j \neq 0$ but small, even if some $p_j < \alpha_0$, it is unlikely any $p_j < \alpha_0/m$.

Bonferroni's method has poor power under Scenario 2.

An alternative procedure follows.

Fisher's combined probability test

The idea is if there are small non-zero effects, it would be more powerful to combine p values rather than just use the minimum p-value.

Suppose, under $H_j : p_j \sim Unif(0,1)$ $H_0 : p_1 \cdots p_m \sim Unif(0,1)$ (The p values here are independent)

Let $x_i = -\log p_i$ (so smaller p_i is, bigger X_i is)

The distribution of X_i,

$$\begin{aligned} Pr(X_i \leq x \mid H_0) &= Pr(-\log p_i \leq x \mid H_0) \\ &= Pr(p_i \geq e^{-x} \mid H_0) \\ &= 1 - e^{-x} \\ &\sim \text{Exp} \end{aligned}$$

Recall, if

$$\begin{aligned} X_i &\sim Exp(1) \\ 2X_i &\sim Exp(2) \\ &\sim \chi^2_2 \\ &\text{So,} \\ -2\log p_j &= 2X_i \sim \chi^2_2 \\ -2\sum_{j=1}^{m} \log p_j &\sim \chi^2_{2m} \end{aligned}$$

Reject H_0 if $-2\sum \log p_j > \chi^2_{2m,1-\alpha_0}$ where α_0 is the target global type I error rate.

Note: Fisher method requires independence.

False Discovery Rate

Suppose you expect a) at least some $\theta_j \neq 0$ and b) at least some null hypothesis will be rejected.

A natural question to ask is, among the θ_Js we declare to be $\neq 0$, what fraction actually are?

This function is called the **false discovery proportion**

Consider the following multiple testing procedure,

- Compute a p-value p_j for each hypothesis $H_j, j = 1 \cdots m$
- Delcare a "discovery" for θ_j for if $p_j < \alpha_E$
- Declare a "null result" for θ_j if $p_j \leq \alpha_E$

Here, α_E is the experimental type I error rate.

Under this procedure, we know that $Pr(D_j = 1 \mid H_j = 0) = \alpha_E$.

And under Bonferroni's procedure,

$$Pr(\text{any } D_j = 1 \mid \text{ all } H_j = 0) \approx m\alpha_E$$

But for large m, controlling these may be irrelevant, but want to know, among the "discoveries", how many are false discoveries.

$$FDP = \frac{\sum_{j=1}^{m} D_j(1-H_j)}{\sum_{j=1}^{m} D_j}$$

Numerator: # $H_{0,j}$ reject but $H_{0,j}$ true and Denominator: # $H_{0,j}$ rejected

The goal is to control the false discovery rate below α, i.e. choose α_E such that,

$$FDR = \mathbb{E}[FDP]$$

Model Assume $H_1 \cdots H_m \sim Binary(\gamma)$

$$\begin{aligned} p_j \mid H_j = 0 &\sim Unif(0,1) = P_0 \\ p_j \mid H_j = 1 &\sim P_1 \end{aligned}$$

Then, $p_j \sim (1-\gamma)P_0 + \gamma P_1$ (the marginal distribution of the p values).

$$\begin{aligned} FDP &= \frac{\sum_{j=1}^{m} D_j(1-H_j)}{\sum_{j=1}^{m} D_j} \\ &= \frac{\sum_{j=1}^{m} \nVdash(p_j < \alpha_E \cap H_j = 0)}{\sum_{j=1}^{m} \nVdash(p_j < \alpha_E)} \\ &= \frac{\sum D_j(1-H_j)}{\sum D_j} \end{aligned}$$

$$\begin{aligned} \mathbb{E}[FDP] &= \mathbb{E}\left[\frac{\sum D_j(1-H_j)}{\sum D_j}\right] \\ &= \mathbb{E}[\mathbb{E}[(...) \mid D]] \\ &= \mathbb{E}\left[\frac{\sum D_j\, \mathbb{E}[(1-H_j) \mid D]}{\sum D_j}\right] \\ &= \mathbb{E}\left[\frac{\{no.D_j = 1\} \times \mathbb{E}[1-H \mid D=1]}{\{no.D_j = 1\}}\right] \\ &= \mathbb{E}[1-H \mid D=1] = Pr(H=0 \mid D=1) \\ Pr(H=0 \mid D=1) &= Pr(H=0 \mid p < \alpha_E) \\ &= \frac{Pr(p < \alpha_E \mid H=0)P(H=0)}{Pr(p < \alpha_E)} \\ &= \frac{\alpha_E(1-\gamma)}{F(\alpha_E)} \\ &< \frac{\alpha_E}{F(\alpha_E)} \end{aligned}$$

The idea is that the FDP is apprximately bounded by $\alpha_E/F(\alpha_E)$.

Definition:

$$FDR = \mathbb{E}[FDP] = \frac{(1-\gamma)\alpha_E}{F(\alpha_E)} < \frac{\alpha_E}{F(\alpha_E)}$$

$$FDR < \frac{\alpha_E}{F(\alpha_E)}$$

FDR control: Choose α_E to be the largest value for which $\alpha/F(\alpha_E) < \alpha$.

Problem: F is unknown CDF of the p-values. F is a mixture of uniformly distributed p-values and p-values from other distribution.