heights and weights, you would to know the distribution of heights for all people with weight 165 lb.

X = height, Y = weight

$X \mid Y = 165$, $E[X \mid Y = y] = \int x f_{X|Y=y}(x|y)\,dx$

$= \sum x f_{X|Y=y}(x|y)$.

Conditional variance and covariance

- The conditional variance is defined as $V(Y|X = x) = \int (y - \mu(x))^2 f_{Y|X}(y|x)dy$, where $\mu(x) = E[Y|X = x]$.
- For random variables X and Y, $V(Y) = E[V(Y|X)] + V(E[Y|X])$.
- For three random variables X, Y and Z, the conditional covariance of X and Y given $Z = z$ is defined as $Cov(X, Y|Z = z) = \int (x - \mu_1(z))(y - \mu_2(z)) f_{X,Y|Z}(x, y|z)dxdy$, where $\mu_1(z) = E[X|Z = z]$ and $\mu_2(z) = E[Y|Z = z]$.
- For random variables X, Y and Z,

$$Cov(X, y) = E[Cov(X, Y|Z)] + Cov(E[X|Z], E[Y|Z]).$$

$$V(Y) = E[V(Y|X)] + V(E[Y|X])$$

$E[Y|X]$ is a r.v. which a fn. of the r.v. X

$V(Y|X)$ is also r.v. which is another fn. of the r.v. X.

$$E[V(Y|X)] + V(E[Y|X]) = V(Y)$$

Quantiles

- To understand the spread of the distribution of a random variable, the concept of quantiles is described.
- Let $0 < \alpha < 1$. The α-th quantile a random variable, denoted by X_α, is given by $P(X < X_\alpha) = \alpha$.
- This is also referred to as the 100α-th percentile.

X, Y, Z, $[X \to$ height, $Y =$ weight, $Z =$ age$]$

How height and weight for a person are correlated at a given age?

age = 60 years. $\quad Z = 60$

$$Cov(X, Y \mid Z=60) = \int (x - \mu_1(z))(y - \mu_2(z)) f_{X,Y|Z}(x,y|z)\, dx\, dy$$

$$\mu_1(z) = E[X \mid Z=60], \quad \mu_2(z) = E[Y \mid Z=60]$$

For $Cov(X,Y) = E[Cov(X,Y|Z)] + Cov(E[X|Z], E[Y|Z])$

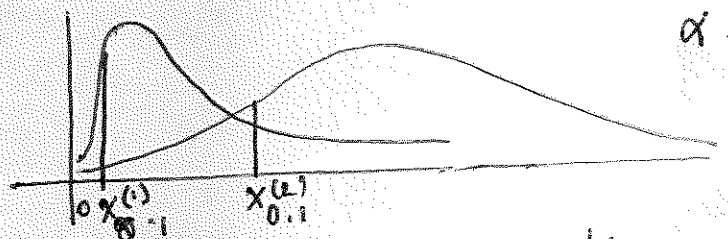
Expectation $\longrightarrow$ average value of a r.v.

Variance $\longrightarrow$ spread of the r.v. around mean.

there is a notion of quantile for a r.v.

a number x_α is known to be the α-th quantile for a r.v. if $P(X < x_\alpha) = \alpha$.

$\alpha = 0.1$



α th quantile is also sometimes referred to as 100α th percentile

$\alpha = 0.1$. $\quad x^{(1)}_{0.1}$ is the 10th percentile for the first r.v.

Moment Generating Functions

- What is the easiest way to find $E[X^k]$ for any k?
- There is a function known as moment generating function given by $M_X(t) = E[e^{tX}] = \int e^{tx} f_X(x)dx$. If *MGF* exists at a neighborhood of 0, then $E[X^k] = \frac{d^k}{dt^k} M_X(t)|_{t=0}$.
- For a random sample, $M_{\bar{X}}(t) = [M_{\bar{X}}(t/n)]^n$.
- **Example:** Let $X \sim N(\mu, \sigma^2)$. Let us compute MGF of X. For every $t \in \mathbb{R}$,

$$E[e^{tX}] = \exp\left(t\mu + \frac{1}{2}t^2\sigma^2\right).$$

- Note that MGF is exists in a range of t. For normal distribution, the range is entire $\mathbb{R}$. However, MGF might not be valid for the entire $\mathbb{R}$ for many other distribution.
- Let $X \sim Gamma(\alpha, \beta)$. Find the MGF of X. (MV moment generating)

$$M_X(t) = E\left[e^{tx}\right]$$
$$= \int e^{tx} f_X(x)\, dx$$
$$= \sum_x e^{tx} f_X(x)$$

If $M_X(t)$ exists in a nbd. of $t=0$

$$E[X^k] = \frac{d^k}{dt^k} M_X(t)\Big|_{t=0}$$

$X \sim Ber(p)$

$X = 0, 1$ w.p. $1-p$ and p

$E[e^{tX}] = 1 \cdot (1-p) + e^t p$ M.G.F

$X \sim Bin(n, p)$

$$E[e^{tX}] = \sum_{x=0}^{n} e^{tx} \binom{n}{x} p^x (1-p)^{n-x}$$

$$= \sum_{x=0}^{n} \binom{n}{x} [pe^t]^x (1-p)^{n-x}$$

$$= [pe^t + 1 - p]^n \quad [\text{by Binomial thm.}] \quad \text{M.G.F.}$$

If $X \sim Bin(n, p)$, then $X = X_1 + \cdots + X_n$

where $X_1, \ldots, X_n \overset{iid}{\sim} Ber(p)$

$$E[e^{tX}] = E[e^{t(X_1 + \cdots + X_n)}]$$

$$= E[e^{tX_1} \cdot e^{tX_2} \cdot \ldots \cdot e^{tX_n}]$$

$$= E[e^{tX_1}] E[e^{tX_2}] \cdots E[e^{tX_n}]$$

$$= (pe^t + 1 - p) \cdots \quad n \text{ times multiplied}$$

$$= (pe^t + 1 - p)^n$$

M.G.F uniquely determines a r.v.

Given a r.v. there is only one M.G.F. associated with it.

$X \sim N(0,1)$

$$E[e^{tx}] = \int_{-\infty}^{\infty} e^{tx} \frac{e^{-x^2/2}}{\sqrt{2\pi}} dx$$

$$= \int_{-\infty}^{\infty} \exp\left[-\frac{1}{2}(x^2 - 2tx)\right] \frac{1}{\sqrt{2\pi}} dx$$

$$= \int_{-\infty}^{\infty} \exp\left[-\frac{1}{2}(x^2 - 2tx + t^2 - t^2)\right] \frac{1}{\sqrt{2\pi}} dx$$

$$= \int_{-\infty}^{\infty} \exp\left[-\frac{1}{2}(x-t)^2\right] \exp\left(-\frac{t^2}{2}\right) \frac{1}{\sqrt{2\pi}} dx$$

$$= \exp\left(-\frac{t^2}{2}\right) \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x-t)^2\right) dx$$

density of $N(t,1)$

$$= \exp\left(-\frac{t^2}{2}\right)$$

$X \sim N(\mu, \sigma^2)$ then, $E[e^{tx}] = e^{\mu t + \frac{1}{2}t^2\sigma^2}$

In normal the m.g.f is defined for all $t \in \mathbb{R}$.

$X \sim \text{Gama}(\alpha, \beta)$

$$E[e^{tx}] = \int_0^{\infty} e^{tx} \frac{1}{\Gamma(\alpha)\beta^\alpha} e^{-x/\beta} x^{\alpha-1} dx$$

$$= \int_0^{\infty} \frac{1}{\Gamma(\alpha)\beta^\alpha} \exp\left(-x\left(\frac{1}{\beta} - t\right)\right) x^{\alpha-1} dx$$

$$= \frac{1}{\beta^\alpha} \frac{1}{\left(\frac{1}{\beta} - t\right)^\alpha} \int_0^{\infty} \frac{\left(\frac{1}{\beta} - t\right)^\alpha}{\Gamma(\alpha)} \exp\left(-x\left(\frac{1}{\beta} - t\right)\right) x^{\alpha-1} dx$$

$$= \frac{1}{(1-\beta t)^\alpha}, \text{ when } t < \frac{1}{\beta}$$

$\text{Gama}\left(\alpha, \frac{1}{\frac{1}{\beta} - t}\right)$ $\quad \frac{1}{\beta} - t > 0 \Leftrightarrow t < \frac{1}{\beta}$

Statistical inference or learning

- So far, we have studied different random variables, their distributions, their associations and spread. Why are the important?
- In statistics, our main job is that given a sample $X_1, ..., X_n$ (lets say a random sample), how to infer on the underlying data generation mechanism.
- Typically, this is done in statistics by assuming $X_1, ..., X_n \sim F$, where F is a distribution and then use the sample to understand F.
- The most common practice in this regard is to choose F represented by only a finite number of parameters.
- For example, depending on the problem in hand, one might consider choosing F as $Ber(p)$ or $N(\mu, \sigma^2)$.
- Clearly, these distributions are represented by only a few parameters and they are called parametric distributions or parametric models.

$X_1, ..., X_n \overset{iid}{\sim} F$

$F \equiv N(\mu, \sigma^2)$

→ parametric models — Gamma(α, β)

Statistical inference or learning

- Sometimes after plotting the data, or by gathering more information on the data collection procedure, one may come to the conclusion that a parametric model is too restrictive to understand the data generation mechanism.
- For example, we may want to keep the distributional form of F unknown, but assume that $f = F'$ and $f \in \mathscr{F}_{SOB}$, where $\mathscr{F}_{SOB} = \{f : \int (f''(x))^2 dx < \infty\}$.
- $\mathscr{F}_{SOB}$ is called the class of Sobolev space of functions, which mainly consists of "smooth" functions.
- Any statistical inference problem can mainly be identified as one of the three types:
  (i) point estimation (ii) estimation of confidence set (iii) hypothesis testing.

estimating a fn. f means estimating $f(x)$ for all x.

$X_1, \ldots, X_n \sim f(\underline{\theta})$

$\theta \in (a_1, b_1)$

Statistical inference or learning

- For the parametric models, point estimation amounts to finding the "best" choice of the parameters fitting the data.
- The confidence set is an interval in which the "true data generating" parameters lie with "high confidence".
- Hypothesis testing amounts testing if data supports a specific hypothesis on the parameters.
- Notions are analogous but little more complicated in nonparametric inference.

$X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ to give point estimate of the density, it is enough to provide ~~point~~ best possible estimate for μ and σ^2.

$X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$ confidence set (a_1, b_1) for μ and (a_2, b_2) for σ^2 are intervals where the true data generating μ and σ^2 respectively lie with high prob.

$\mu = 0$ vs. $\mu \neq 0$.

Parametric model $\longrightarrow$ point estimation

Parametric Inference

- In parametric modeling, we assume that F belongs to $\{f(x;\theta) : \theta \in \Theta\}$, where $\Theta \in \mathbb{R}^k$ is the parameter space and $\theta = (\theta_1, ..., \theta_k)'$ is the parameter.
- The problem of inference then reduces to the problem of estimating the parameter θ.
- How would we ever know that the distribution that generated the data is in some parametric model?
- Indeed, we would rarely have such knowledge which is why nonparametric methods are preferable.
- Studying parametric models is useful for two reasons.
- First, there are some problems where a parametric model seems not a bad fit.
- Secondly, it helps us to provide background for understanding certain nonparametric models.
- For parametric modeling, we will discuss two ways of providing the point estimation: (i) method of moment (ii) maximum likelihood estimator.

$X_1, \ldots, X_n \overset{iid}{\sim} F$

Method of Moments Estimator

- Let $m_j = \frac{1}{n}\sum_{i=1}^{n} X_i^j$
- Let $\mu_j = E[X^j]$. Generally μ_j's are functions of the unknown parameters $\theta_1, ..., \theta_k$. Therefore by solving k equations

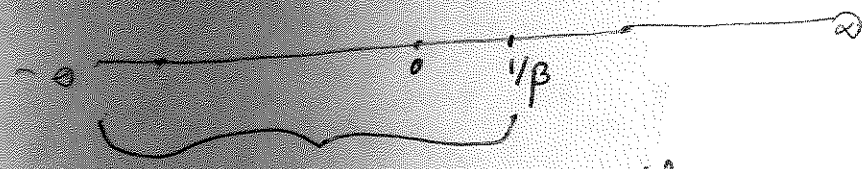
$$m_j = \mu_j'(\theta_1, ..., \theta_k), j = 1, ..., k,$$

  we have some estimates of $\theta_1, ..., \theta_k$. They are called method of moments (MOM) estimates.
- **Example:** $X_1, ..., X_n \sim N(\mu, \sigma^2)$. MOM does the following

$$\bar{X} = \mu, \ \frac{1}{n}\sum_{i=1}^{n} X_i^2 = \mu^2 + \sigma^2 \Rightarrow \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

- Here MOM estimator are "good" estimators.
- However, consider the situation when $X_1, ..., X_n \sim DE(\mu, \sigma^2)$. Even for this case MOM estimator remains the same.
- This is a disadvantage that the MOM estimator doesn't take care of the difference in distributions.

$\frac{1}{(1-\beta t)^\alpha}$ $t < \frac{1}{\beta}$ remember $\beta > 0$



Let's say $X_1, \dots, X_n \overset{iid}{\sim} f(x, \underline{\theta})$ $\underline{\theta} = (\theta_1, \dots, \theta_K)'$

$E[X^K] = \frac{1}{n}\sum_{i=1}^{n} x_i^K$ $E[X^K] = \int x^K f(x,\theta)\, dx$

$m_K = \frac{1}{n}\sum_{i=1}^{n} x_i^K$

$\mu_K = E[X^K]$

$\left.\begin{aligned} m_1 &= \mu_1 \\ m_2 &= \mu_2 \\ &\vdots \\ m_K &= \mu_K \end{aligned}\right\}$ K equations

Example: $X_1, \dots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$

I am trying MOM estimators for μ and σ^2.

$E[X] = \frac{1}{n}\sum_{i=1}^{n} x_i$ - - - (1)

$E[X^2] = \frac{1}{n}\sum_{i=1}^{n} x_i^2$ - - - (2)

$E[X] = \mu$, $E[X^2] = \mu^2 + \sigma^2$ $Var(X) = E[X^2] - (E[X])^2$

$\sigma^2 = E[X^2] - \mu^2$

$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$ - - - (1)

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

$\hat{\mu}_{MOM} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$, $\hat{\sigma}^2_{MOM} = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$

(a) $DE(\mu, \sigma^2)$ has the form $\frac{1}{2\sigma} \exp\left(-\frac{|x-\mu|}{\sigma}\right)$, $-\infty < x < \infty$

$E[X] = \mu$ and $Var(X) = \sigma^2$

$X_1, \ldots, X_n \overset{iid}{\sim} DE(\mu, \sigma^2)$

$\frac{1}{n}\sum_{i=1}^{n} X_i = E[X] = \mu$

$\frac{1}{n}\sum_{i=1}^{n} X_i^2 = E[X^2] = \mu^2 + \sigma^2$